Altering speech synthesis prosody through real time natural gestural control

David Abelman

Master of Science

Artificial Intelligence

School of Informatics

University of Edinburgh

2013

Abstract

A significant amount of research has been and continues to be undertaken into generating *expressive prosody* within speech synthesis. Separately, recent developments in HMM-based synthesis (specifically pHTS, developed at University of Mons) provide a platform for *reactive* speech synthesis, able to react in real time to surroundings or user interaction.

Considering both of these elements, this project explores whether it is possible to generate superior prosody in a speech synthesis system, using natural gestural controls, in real time. Building on a previous piece of work undertaken at The University of Edinburgh, a system is constructed in which a user may apply a variety of prosodic effects in real time through natural gestures, recognised by a Microsoft Kinect sensor. Gestures are recognised and prosodic adjustments made through a series of hand-crafted rules (based on data gathered from preliminary experiments), though machine learning techniques are also considered within this project and recommended for future iterations of the work.

Two sets of formal experiments are implemented, both of which suggest that - under further development - the system developed may work successfully in a real world environment. Firstly, user tests show that subjects can learn to control the device successfully, adding prosodic effects to the intended words in the majority of cases with practice. Results are likely to improve further as buffering issues are resolved. Secondly, listening tests show that the prosodic effects currently implemented significantly increase perceived naturalness, and in some cases are able to alter the semantic perception of a sentence in an intended way.

Alongside this paper, a demonstration video of the project may be found on the accompanying CD, or online at ***http://tinyurl.com/msc-synthesis***. The reader is advised to view this demonstration, as a way of understanding how the system functions and sounds in action.

Acknowledgements

Thank you to my supervisor Rob Clark for his input and assistance throughout the course of this project, the staff and fellow students who have provided useful suggestions along the way, and the friends who kindly donated their time to help me as part of the evaluation phase. Finally, a big thank you to all family, friends and loved ones who have supported me over the years. I really appreciate everything.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*David Abelman*)

Table of Contents

Chapter 1

Introduction

Chapter summary: *This chapter introduces the problem statement, provides a brief history for context, and outlines motivations for carrying out the work.*

1.1 Problem statement

The aim of this project is to answer the question of whether it is possible to **generate superior prosody in a speech synthesis system, using natural gestural controls, in real time**. This primary problem statement is split into various sub-parts that are discussed, tested and evaluated within this project.

To illustrate the proposed system with a few examples, potential tasks for such a speech synthesis system may be to:

- emphasise important words in a statement as a user gesticulates with their arms
- synthesise a statement with interrogative prosody as a user shrugs their shoulders
- articulate words more clearly if a user's body language suggests the spoken information is significant

If successful, future evolutions of the work should extend to a wider number of gestures and prosodic effects, as well as being able to learn customised gesture to prosody mappings from individual users.

1.2 Background and context

The field of speech synthesis has engaged and challenged scientists for hundreds of years, though it is the last 50 years in particular that the technology has undergone significant and regular breakthroughs [1]. **Concatenative synthesis**, developed throughout the second half of the 20th century, operates under the idea that realistic sounding synthetic speech can be generated by piecing together small units of real recorded speech. Early implementations involved the splicing of magnetic tape, with each separate unit corresponding to a phone [2]. **Diphone synthesis** - where each individual unit consists of a diphone extracted from a 'carrier sentence' - followed as a mainstream research topic through the 1980s. The concatenated diphones would undergo signal processing (TD-PSOLA) in order to 'mould' the prosody to that predicted by a simple prosody model [3] [4].

Unit selection synthesis came as a breakthrough in the late 1980s as methods shifted from using singly stored examples of each diphone, to a database storing many different versions of each diphone [1]. Given a choice of multiple diphones to use for any one unit of synthesised speech, cost functions for multiple combinations must be calculated, and the diphone sequence incurring the minimum overall cost is used. The cost function consists of a 'target cost' (linguistic features such as phonetic context, prosodic context and syllable position) and 'join cost' (acoustic features such as cepstral distance and frequency). The Viterbi algorithm and pruning methods are used to efficiently establish an optimal sequence of units to concatenate [4] [5].

In recent years, **statistical parametric models of speech synthesis** - the type of synthesis utilised within this project - have become competitive with the concatenative methods outlined. In contrast to unit selection, statistical methods do not select actual recorded units of speech in order to synthesise. Output waveforms are generated from parameters stored within the model, with these parameters essentially acting as 'averages' calculated from a recorded speech database. Although the quality of the very best examples of unit selection synthesisers may still be argued to be superior to the very best statistical methods [6], statistical methods do boast various advantages over concatenative methods. These centre around the flexibility that a statistical parametric method offers. Voice characteristics, accent and emotional feeling can all be flexibly altered by adjusting model parameters, with only a limited amount of data required to do so [6] [4]. HMM-based synthesis (the statistical parametric model used within this work) uses hidden Markov models as the generative models in question. More

technical detail on the workings of HMM-based synthesis are provided in Section 2.

However, despite the significant advances in speech technology in recent years, much work remains to improve synthesisers to a state where they are comparable to human voices in any situation. Specifically within HMM-based synthesis, research currently spans a wide range of areas including emotional speech synthesis, expressive prosody, voice quality, interpolation between accents and styles, and real-time control. This project focuses on two particular areas and their interaction: **expressive prosody** and **real-time control** of the synthesis. The aim is to use recent developments in real-time control of speech synthesis to alter the prosody of sentences, as directed by a user, resulting in more natural and expressive speech.

1.3 Motivation and potential applications

Producing reactively expressive speech is a significant challenge facing speech synthesis today, though one on which there has been little research to date [7] [8]. A small number of papers on the subject are outlined in Section 2.2, though as of yet there are relatively few practical applications in development.

Although the system created in the course of this project will act primarily as a pilot, a fully functional system would have various potential applications. A primary use may be in text-to-speech communication aids of those with vocal disorders. More natural expression and prosody may be 'conducted' by the user in real-time, either through a set of 'standard' natural gestures, or through a set of custom-designed gestures for those with physical disabilities (for example, eyebrow or finger movements).

Additionally, technology developed as part of this system may be incorporated within potential 'sign-language synthesis' systems of the future [9] [10]. In addition to synthesising words based on sign-language hand movements, the manner in which the gestures are performed may indicate to the system a certain expressive or emphatic style in which to synthesise the speech.

Other potential applications may exist within the entertainment industry. For example, the technology may be adapted for use within synthesised singing voices, or perhaps within future 'instrument-voice hybrids' that people may wish to control through body gestures. Ultimately, any situation in which it would be useful to improve expressiveness of a voice-like synthesis in real-time would benefit from the research that this project intends to undertake.

Chapter 2

Previous work and literature

Chapter summary: *This chapter introduces and critiques previous work in the following areas: HMM and reactive HMM synthesis, expressive synthesis, selected prosodic effects, gestures within speech, and basic gesture recognition using the Microsoft Kinect.*

2.1 HMM synthesis overview

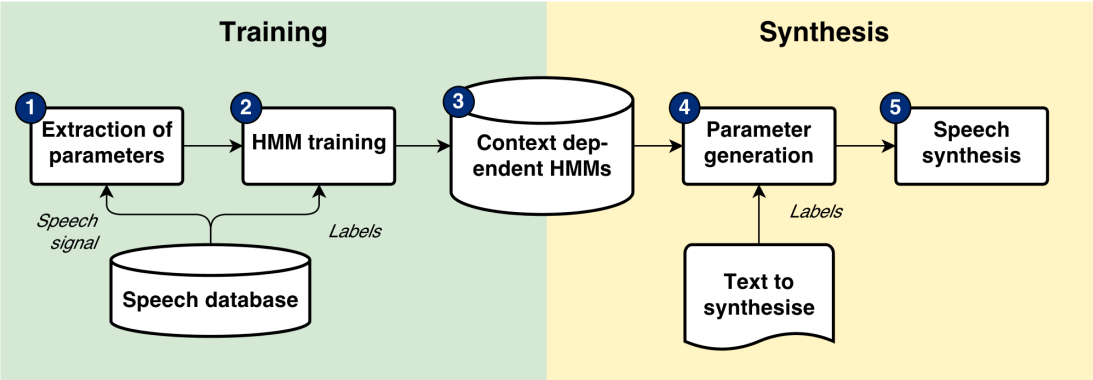
The aim of this section is to provide context for the following discussion of pHTS. As such, a limited number of papers are used to outline the primary mechanics of HMM-based synthesis.

HMM Speech Synthesis Systems (HTS) are a particular statistical parametric model of speech synthesis, using hidden Markov models in the generation of speech parameters. As previously noted, the method generates speech parameter trajectories using model parameters trained on multiple examples of speech within a recorded database, rather than selecting stored examples to 'replay' (as in the case of unit selection) [11].

It is simplest to consider such systems in two parts, which share a broad symmetry: the training phase, and the synthesis phase, as shown in Figure 2.1. Sub-processes are now outlined within these, based primarily on the descriptions in [6], [12], and the primer in [11].

1 - Extraction of parameters Parametric representations of real speech are extracted, using vocoder technology. These parameters describe properties such as F_0, and the spectral envelope of the speech.

Figure 2.1: *Basic HMM-based synthesis schematic. Based on more detailed diagram to be found in [6].*



2 - HMM training The extracted parameters are modelled in a hidden Markov model framework, taking into account linguistic, phonetic and prosodic contexts. However, as a context may span a whole utterance, very few units share identical contexts. Thus contexts in the training data are *clustered*, using a decision tree framework. The important point to note here is that as the contexts used span the entire sentence, traditional HTS methods cannot be *reactive* in nature, as the models for each unit are based on the context for the whole sentence.

3 - Context-dependent HMMs stored Spectrum, excitations and durations are all stored within the HMM framework, ready for synthesis.

4 - Parameter generation Input text is converted to a set of context dependent labels. Speech parameters are generated using the stored HMM data, with the output set of parameters being determined by maximum likelihood. Once again, this likelihood calculation is maximal based on using the whole sentence as context, thus the synthesis cannot be *reactive* in nature. Note that a naïve method generates a piecewise stationary set of parameters. In reality, parameter time derivatives, and double derivatives, are used to generate smooth parameter trajectories.

5 - Speech synthesis The speech waveform is synthesised from the parameters generated above.

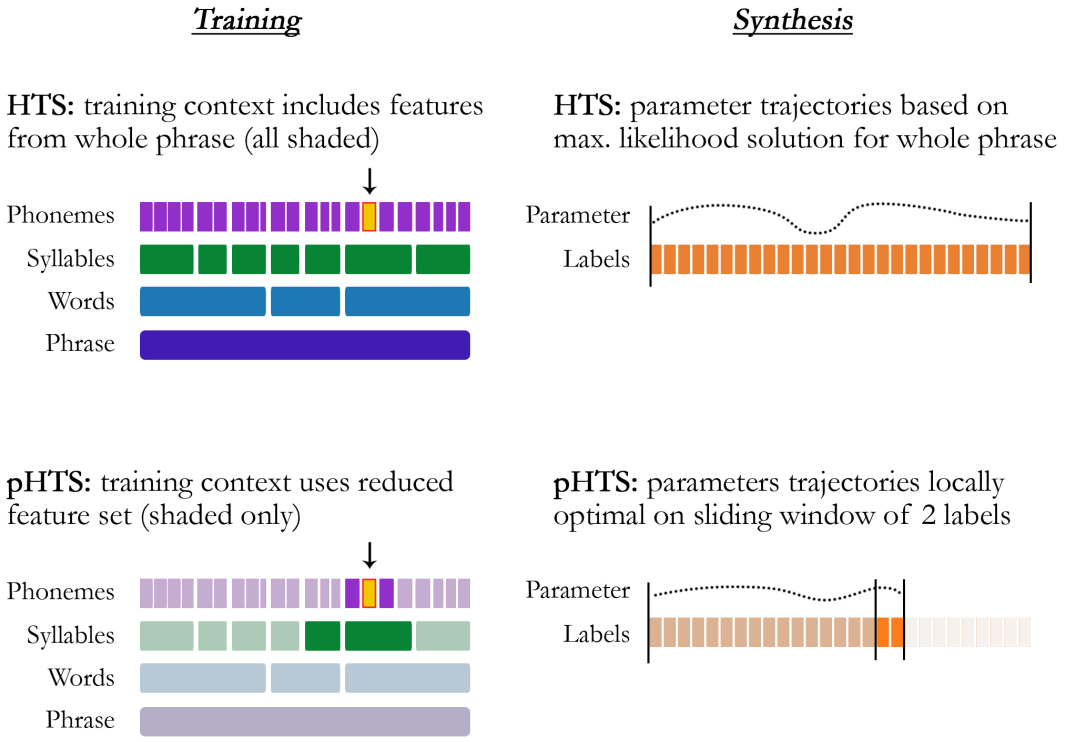
2.2 Performative HTS (pHTS)

In the modern era, research into the modification of synthesis output *in real time* has been very limited. One group at University of Mons, however, has recently undertaken a significant amount of work in this area [7]. Under the belief that converting text to speech at a sentence level restricts the range of potential applications, the group have developed an alternative synthesis engine that allows waveforms to be generated in real time, with the synthesiser mid-utterance. Applications for the technology are claimed to fall into two primary areas: ***context reactive speech synthesis*** (in which synthesised text reactively adapts according to the surrounding conditions) and ***performative speech synthesis*** (in which synthesis may be made more expressive under a user's control) [7]. This project focuses on the latter, although it would also be possible to incorporate elements of reactivity to the surrounding environment in future.

The modified HTS engine developed is called pHTS, standing for ***performative*** HTS. In order to make the system reactive, the phonetic context required in calculating the synthesis parameters is reduced from that of the whole sentence to that of a much smaller window. This change requires two main modifications. Firstly, the context used in training the model is reduced to just the current and surrounding phonemes, and the current and previous syllable. This is in contrast to standard HTS, where features from the whole utterance are considered. Secondly, during synthesis, the generation of parameters occurs on a sliding window of two labels. This means that likelihood maximisation occurs based on the concatenation of just two HMMs (those of the current and previous label), as opposed to HMMs concatenated for the whole sentence, as is the case with standard HTS. The parameter trajectories for the whole phrase are therefore not maximally probable globally, but are maximally probable locally [7]. These changes to the training and synthesis are summarised in Figure 2.2.

The group have carried out both objective and subjective tests on the modified engine in both [13] and [7], with comparable results. Outlined here is the evaluation from [7], virtue of being more recent and in-depth. Standard HTS is compared with two versions of pHTS. Objective tests compared mel-cepstral distortion (Mel-CD) and root-mean-square error in F_0 (RMS F_0). With regards to Mel-CD, it was found that both of the pHTS systems introduced around 1dB of distortion relative to HTS, which the authors claim is 'very close', on the basis that 1dB is usually considered the smallest noticeable threshold for spectral distortion. Meanwhile, RMS F_0 error in voiced regions of synthesised text ranges sits between 80 and 100 cents for the pHTS sys-

Figure 2.2: *Diagram illustrating basic differences in training/synthesis for HTS/pHTS.*



tems. Again, considering that 25 cents is the threshold for noticeable pitch differences (for pure tones), 100 cents is considered to be small by the authors. Subjective tests showed that listeners tended to rate HTS between 0.1 and 0.7 higher than pHTS on a 7 point scale of 'quality' (at 95% confidence). This is considered to be 'relatively minor' by the authors. Further experiments are carried out into interpolation between hypoarticulated and hyperarticulated speech, which are not expanded on here.

Having been able to compare HTS and pHTS directly throughout the course of this MSc project, the author believes that the quality difference presented by the group in [7] may be slightly understated. It often appears to be the case that certain sounds are synthesised reasonably poorly in pHTS, when no such issue occurs with HTS. The difference in subjective rating of 0.1 to 0.7 is not compared to any similar ratings within [7], so it could be argued that this difference may not necessarily be 'relatively minor' as claimed by the original authors. The work presented in this MSc project ultimately does not compare any output with standard HTS (the baseline audio files all use neutral pHTS synthesis), and based on the work in [7] it is assumed HTS and pHTS are of similar quality. However, this would benefit from independent investigation. It

should also be noted that this project uses code based on MAGE 1.00, whereas an updated version (MAGE 2.0) is now available.

A variety of potential applications for pHTS have been outlined and developed by the group. These include *HandSketch*, a pen-based musical instrument prototype [13], *CoVop*, a multi-user 'social game' allowing different users to control different aspects of a synthesiser [14], speech synthesis based on face-tracking [15], and accent interpolation through an interactive map application [16]. The design of systems that reactively respond to their surroundings (such as interpolating to more hyperarticulated speech as background noise levels increase) is also discussed [7].

Finally, [8] incorporates skeleton tracking (using Microsoft Kinect) into the pHTS system to create a reactive speech synthesiser, in which pitch and duration are controlled by hand movements. It is found that meaningful expressiveness is difficult to simulate when pitch and duration modulations are mapped directly to the spatial coordinates of the hands in this particular way. It is this work that this project intends to build on.

2.3 Expressive Synthesis

Although synthetic speech has increased in both naturalness and intelligibility to unprecedented standards, the ability to add realistic *expression* to synthesised voices remains in general less satisfactory. It is however an area in which there has been a considerable amount of work. Some formant synthesisers of the 1990s benefited from acoustic-based 'emotional modification' modules, designed according to hand-crafted rules modifying pitch, rate of speech, voice quality and articulation [17], as in the case of [18] and [19]. Within diphone synthesis, research has attempted to add expression both through signal processing [17], and through recording the actual diphone database in various styles, as in [20].

Expressivity itself is a broad concept, encompassing emotional speech, focus and emphasis, styles of speaking, and more [21]. The following outlines more recent research efforts both in terms of ***emotional expressivity***, and in modelling ***expressive prosody*** (and in particular, emphasis). The section is split according to unit selection and HMM-based synthesis.

2.3.1 In unit selection

Emotional expressivity As outlined by [17], the nature of unit selection does not generally lend itself to signal processing methods. Therefore attempts to incorporate emotion into unit selection systems usually proceed by creating recordings in a variety of specified styles. Depending on the style of synthesis required, the units are then selected from the appropriate database at synthesis time. Work such as [22] and [23] follow this methodology, recording databases in happy/sad/angry and shouted/spoken styles respectively.

Expressive prosody Basic unit selection synthesisers are generally good at reading neutral, 'newspaper-style' text, but typically produce poor results when it comes to synthesising prosody indicating some specific meaning - for example, converting a statement into a question, or emphasising certain elements in an utterance [24]. Prosodic cues such as these occur commonly in human speech and are integral to proper information delivery [25]. Various attempts to address this involve work at the *script design* level, so that appropriate prosody exists within the database from which units are selected for synthesis. [26] designed a 'sub-database' to be recorded by a voice talent nine times, in a variety of styles (combinations of fast, high, slow and low), in order to increase a system's prosody range, without experiencing the quality deterioration that would come with signal processing. More recently, [24] and [27] develop an improved recording script for 'emphatic accents' and phrase boundaries (to allow more realistic question-type prosody). The script consists primarily of word lists (to be read with varying intonation) and Lewis Carroll literature. In the first test, emphasis was recognised 40% of the time (chance level = 18%). In the second test on the system with the phrase boundary component, the system was preferred 56% of the time (chance level = 50%).

In contrast to methods centred around script design, it is also possible to use a 'prosodic phonology' approach, as described in [28]. Here, 'statistical acoustic-prosodic models' are built, linking specific prosodic units to predictable changes in signal parameters. ToBI pitch accents are used to mark prosody, and F_0 and duration models are trained on the basis of these markings. It was found that in the case of contrastive emphasis, words were typically marked with pitch accent H* and phrase accent L- (in ToBI notation), with the pitch accent corresponding to a 28% increase in pitch on average. A speech synthesiser using these rules as cost function features produced contrastive emphasis that was correctly identified by listeners over 80% of

the time. The subjective evaluation methods used in this paper are similar to those that this project intends to employ (see Section 5).

2.3.2 In HMM-based synthesis

Emotional expressivity As with unit selection, it is possible to record databases in different emotional speaking styles, upon which the HMMs are subsequently trained. For example, [29] records four speaking styles, showing that 80% of speech samples are correctly classified into the correct style by listeners. Additionally however, statistical parametric methods such as HTS have the advantage by which a voice may be *adapted* to another style by shifting its parameters appropriately. This provides the advantage of not necessarily requiring a full set of training data for every emotion or style to be synthesised. Only a small amount of data is required to draw up parameter adjustments that can be applied to a neutral or primary database. This technique of *'style adaptation'* is proposed in [30].

Expressive prosody Some of the most significant work into realising contrastive emphasis within HMM-based synthesis has been carried out at the Centre for Speech Technology Research in Edinburgh, such as detailed in [31]. Emphasis is synthesised through modifications in the synthesiser's context dependent labels (pitch accents are marked on surrounding/current syllables and words, and emphasis marked on surrounding/current phonemes and syllables). The contrastive word pairs are marked up automatically, using textual features only. However, subjective listening tests showed an overall preference for the *non-emphatically* synthesised sentences, both in the case of textually 'contrastive' and 'non-contrastive' sentences. The hypothesised explanation for these results was that the generated emphasis was often stronger than appropriate.

A subsequent piece of work by the same authors [32] compared the results of a binary 'accent vs. non-accent' synthesiser to one using three types of pitch accent (the third conveying 'contrastive focus'). The model using three levels of accent was preferred in significantly more cases than the model using two. This suggests that two levels of pitch accent within a synthesiser may be flattening a hierarchical structure of prosodic prominence in too simplistic a way [32].

Various other studies focus on emphatic HMM-based synthesis. For example, [33] built two separate synthesisers using emphasised speech data. The first trains two

separate HMM sets at training time - one using emphatic speech, and one using non-emphatic speech. In the second, a mixed model is trained, using both the emphatic and non-emphatic speech data simultaneously. This second model is found to sound more natural, though the degree of emphasis is judged to be slightly lower [33].

Work has also been carried out using *natural* speech recordings, as opposed to specially collected emphatic speech recordings. These contain weaker emphatic clues, which are more affected by suprasegmental features of the speech. [34] introduces two alternative clustering methods (both decision-tree based) to separate emphatic and non-emphatic units, where traditional clustering methods fail. Meanwhile, [35] focuses on labelling emphasis in normal speech in an unsupervised manner. This is accomplished by tracking the difference in F_0 between synthesised and real speech. When this difference passes a certain threshold, the real speech is flagged as being emphasised. This data is then used to build an emphatic HMM-based synthesiser, whose performance is comparable to one trained with manually labelled emphatic speech.

2.4 Prosody of speech

The study of prosody (not restricted to that within speech synthesis) is an extensive field, and as such this section introduces a limited number of areas of interest, given the work carried out in this project.

2.4.1 Definitions

A whole host of terms are used to describe a set of interrelated concepts within the literature, and it is useful to add clarity with some initial definitions:

Stress A syllable is either stressed or unstressed (though it is often argued that there are in fact three or more levels of stress in English) according to the amount of effort expended in speaking the syllable [36]. This may be reflected in pitch prominence, length, roundness of vowels, spectral tilt, or a combination of these. According to [37], each prosodic constituent (known as a 'foot') will contain a stronger syllable, which is considered to be stressed. A common example used to demonstrate stress on different syllables is to consider the word *permit*, in both its noun (*PERmit*) and verb (*perMIT*) forms. Stressed syllables have the *potential* to be pitch accented.

Pitch accent A stressed syllable which is marked specifically by a change (often a rise) in F_0 [37].

Prominence A syllable standing out from the point of view of the listener (due to general stress or pitch accent) is considered to be a prominent syllable [36].

Emphasis The use of prominence to demonstrate the importance of a word or concept [36].

Focus A speaker will put extra effort into a part of the sentence considered to be most significant; this is known as the focus of a sentence. According to [38], focus can be split into 'broad' and 'narrow' focus. Broad focus refers to expression-wide focus, for example *'I didn't give him a sandwich, I gave him five francs'*, whereas narrow focus belongs to a smaller constituent, such as *'five'* in *'I didn't give him three francs, I gave him FIVE francs'* [38].

Nuclear accenting, prenuclear accenting, postnuclear unaccented The final accent in a phrase is referred to as a nuclear accent, and is regarded as the most prominent accent in a phrase. Specifically unaccented words following a nuclear accent are referred to as being postnuclear unaccented. Any accents preceding the nuclear accent are referred to as being prenuclear accents [39]. For example, consider the response to *'What did the girl admire from a distance?'*. Within *'The **girl** admired the **CANYON** from a distance'* we see a prenuclear accent (***girl***), a nuclear accent (***CANYON***) and a postnuclear unaccent (*distance*) [39].

ToBI notation A guide to ToBI notation (a standard for describing prosody) is not included in this report for brevity, though a particularly informative introduction, including audio clips, can be found on Macquarie University's Department of Linguistics Phonetics and Phonology course page[1]. The underlying tonal theory is that of Pierrehumbert, found in [40]. This report primarily references pitch accents, phrase boundaries, and boundary tones.

[1] http://clas.mq.edu.au/phonetics/phonology/intonation/tobi_introduction.html

2.4.2 Prominence through pitch accents

In a sentence, some words will be more prominent than others, courtesy of being marked by a pitch accent. These words will be more salient to a listener. In addition to the altered F_0, the word is also often made more salient through changes in volume and duration [3].

High pitched words, with more energy put in their production, are more easily heard by the human ear. Thus for efficiency, items most worthy of a listener's attention (for example, new information in the discourse) tend to be pitch accented by the speaker, to reduce the cognitive load of the listener as much as possible [41]. According to Pierrehumbert, any high pitch marking (containing H*) represents new information in the discourse, whereas low pitch accents (containing L*) are used to highlight concepts already presupposed within the context [41]. This project aims to explore both H* and L* pitch accents in some form.

Various studies have attempted to automate the recognition of prominence. For example [42] uses conditional random field (CRF) models to detect prominence from acoustic features. Interestingly, the features with highest information gain were the duration of the word, and the *standard deviation* (i.e. variability) of pitch and energy across prominent words, rather than their absolute pitch. For words of more than one syllable this makes sense, as a pitch accent adds much variability in pitch on the stressed syllable, relative to the rest of the word. For one syllable words, this suggests that a pitch accent is not a simple pitch raise, but pitch *variation* - for example, a raised pitch initially that falls over the course of the syllable.

2.4.3 Contrastive emphasis

We now move from general prominence, to the prominence caused by some 'contrastive' concept. The notion of contrast itself has long been debated - one fundamental question argued was whether the concept of ordinary focus should encompass the concept of contrastive focus. Chomsky believed that any nuclear accent not obeying the Nuclear Stress Rule (which states that the nuclear stress must fall on on the last word able to carry an accent) indicates *contrastive focus*. In opposition, Bolinger does not define contrastive focus as existing at all in its own right - any focus whatsoever will establish some set of possible alternatives, and is therefore in some sense contrastive. As potential alternatives are narrowed down, the pitch accent approaches what is thought of as a contrastive accent [43].

Rather than becoming embroiled in the definition of 'contrastive', this study is more interested in the prosodic effect of emphasising words in a contrastive manner. The prosodic aspects primarily affected are pitch, duration, and spectral intensity. In [39] we see that duration increases for more prominent accents (with the most prominent type of accent being contrastive), as does the spectral intensity. In terms of pitch, it is generally the L+H* pitch accent that is associated with contrastive emphasis [44]. Within [39] this is followed by a low phrase accent and boundary tone (L-L%). Similarly, a number of significant researchers in the field have all identified contrastive topics as being associated with a pitch accent L+H* (Pierrehumbert, Steedman, Gundel, Fretheim, as outlined in [44]). It is this way in which this project would ideally intend to consider contrastive emphasis (L+H* L-L%).

2.4.4 Questions

There are number of different question types in English, including Yes/No questions (*Are you coming to the cinema?*), WH-questions (*What shall we watch?*), alternative questions (*Would you rather watch Titanic or Love Actually?*) and tag questions (*You haven't seen Titanic, right?*). Although a considerable amount of work has addressed both the semantics and pragmatics of questions (for example [45] and [46]), relatively little literature exists on the prosody of questions in English [47].

However, the dimension of pitch is addressed within [47], which analysed a set of natural speech taken from television broadcasts for Yes/No and WH-question intonation, following ToBI conventions. Three elements of the sentence prosody are explored: the 'locus of interrogation' (i.e. the pitch accent on the WH-word, or the fronted auxiliary in yes/no questions), the 'nuclear tune' (i.e. the intonation at the end of questions), and the 'topic pitch accent' (i.e. the pitch accent used on the topic of the question).

In terms of locus of interrogation, it was found that the WH-word was marked with a L+H* or H* accent in the vast majority of cases. In contrast, for yes/no questions the fronted auxiliary was marked with no pitch accent or a low accent (L*) about as often as a high accent (H*). With regards to nuclear tune, WH-questions tended to end in a fall (L-L%), whilst yes/no questions were more evenly split between falling (L-L%) and rising (H-H%). Finally, in terms of pitch accent on the topic of the sentence, there is a great deal of variability, with a slight preference for accents of the H* / L+H* kind. Conversely, the remainder ($\sim 30\%$) of sentences were deaccented, or marked with L*.

As the overall number of sentences analysed is low, the results must be taken as directional in nature. Indeed, a later study by the same author [48] looks in more detail at the tonal constituents of yes/no questions specifically. In this case, nuclear tune was found to be marked primarily by a low rise (L* H-H%), with fewer examples of falling (L-L%) endings observed. In addition, more low heads (L* accents on the topic of the sentence) were observed in this study, and are something that this project will attempt to model.

Another study by the author [49] focuses on WH-question prosody in further detail. Many previous accounts have explained the falling nature of WH-questions (L-L%), either by the fact that the question pre-supposes some information (explaining the low phrase boundary) [50], or that a WH-question has a more demanding nature [51]. This study confirms that high falls (H* L-L%) and rise falls (L+H* L-L%) provide the majority of WH-question prosodic contours.

2.5 Gestures within speech

Spoken language and physical gesture and closely interrelated - it has been said that about 90% of descriptive speech is accompanied by some sort of gesture [52]. There are three primary gesture types: *deictic gestures* - those that point towards objects referred to within the speech, *iconic gestures* - those that attempt to describe some physical property of the object being referred to, and *beat gestures* - those that do not contain semantic content, but mark important units or sections of the speech [53]. This project will focus primarily on beat gestures, which include flicks of the hand, eyebrow movements, and head movements [54].

Most relevant for this project is the relative timing of beat gestures with regards to pitch accents. Various studies have addressed this area. For instance, [55] finds that the apex of a beat gesture is centred on the spoken pitch accent. The data is normally distributed around this point, with a standard deviation of $\sim$300ms. Separately, [56] presents similar findings with regards to the apex of the gesture aligning with the pitch accent, though also tracks other anchor points within the motion. Their tracking methods influence some of the preliminary experiments in this work (see Section 3.2.4). Additionally, this study also explores peoples' perceptions of beat gestures that are shifted relative to the speech. It is found that subjects are relatively insensitive to beat gestures performed early, but very sensitive to gestures performed late relative to the pitch accent. This will be beneficial from the perspective of this project - an earlier

gesture will allow more leeway in terms of latency.

2.6 Gesture recognition

The automatic recognition of gestures from video footage has been studied for many years. However, the launch of Microsoft's Kinect in late 2010, with its raw depth sensing capabilities, has altered the human-computer interface landscape. Whereas extracting features from 2D video data is computationally expensive, and thus challenging to use in real-time applications, the Kinect SDK includes code to build a model of a user's skeleton in real time [57]. Little of the academic literature on gesture tracking with Kinect actually uses this *skeletal* data - for example [58] discuss algorithms to use the depth camera's raw data to remove background from consecutive frames, and recognise gestures based on frame by frame differences. However, [57] *does* discuss using the readily available skeletal data for gesture recognition. The paper uses a nearest neighbour classifier to recognise eight different hand gestures, with 99% accuracy. Four joints (both hands and both elbows) are tracked, relative to the spine joint, which is used as a reference. This MSc project ultimately employs a similar joint tracking method (hand movements in relation to a hip reference), though instead of classifying gestures using machine learning techniques, simple rules based on joint coordinates trigger the recognition of gestures. A useful extension to this project would be to apply the more robust classification methods used in [57] to this work.

Chapter 3

Design

Chapter summary: *This chapter steps through a number of design choices made, outlines some basic preliminary experiments carried out to establish system parameters (such as pitch/duration shifts, gesture recognition rules), before describing the system implementation, and issues encountered.*

3.1 Choices and rationales

A number of broad design choices have been considered prior to building the mechanics of the system. The primary options are outlined here, along with eventual decisions made and their rationales.

3.1.1 Pre-sentence control vs. live gestures

Given the need to modify prosody using some motion sensing input device, two possible system formats have been considered:

1. **Pre-sentence gestures:** A gesture would be performed prior to the sentence being uttered, which would indicate a certain type of prosodic effect or structure to be applied to the synthesis - for example, 'a declarative sentence with a nuclear accent on the first noun'.

2. **Live gestures:** The gestures would be performed at the same time as the synthesised text is uttered. These would be able to control both specific effects on certain syllables or words (for example, emphasis on a noun or an immediately preceding adjective), or indicate certain sentence structures as described above.

Decision: As previously discussed, this project aims to utilise pHTS in modifying the prosodic form of sentences whilst they are being spoken, therefore using ***live gestures*** will be the most appropriate and interesting option to pursue. Additionally, this will allow for more specific 'word by word' control, the synthesiser will not be forced to pause between sentences whilst the user gestures, and the user may alter their intended prosody even whilst the text is being spoken.

3.1.2 Choice of prosodic effects

A huge choice of prosodic effects may be implemented. The following outlines the effects chosen for this prototype system; factors considered include ease-of-coding, how frequently such effects occur in speech, and the inclusion of an interesting variety of intonations.

- **'Contrastive emphasis':** Based on the ToBI tones L+H* L-L%, this type of emphasis acts as a nuclear accent in a phrase. Any text subsequent to the accent is unemphasised (i.e. treated as pre-supposed background information).

- **'General emphasis':** Based on the ToBI tone H*, this type of emphasis can be used as a general pitch accent in a phrase, to add prominence at a lesser level than contrastive emphasis (as described in Section 2.4.2). This type of accent does not unemphasise any of the following text, acting as a prenuclear accent.

- **'Yes/No question':** As alluded to within Section 2.4.4, a Yes/No question may take many prosodic forms, though some are more common than others. This prototype will aim to model two forms, one with a high head (more common) and one with a low head (less common). Both will end in a high boundary tone (H%).

- **'WH-question':** Likewise, various prosodic forms exist for WH-questions. This project will aim to model the most common type, namely the high fall (H* L-L%) discussed in [49].

- **'Extended periods of 'important' and 'unimportant' speech':** In addition to the specific contours of prosody outlined above, it will also be interesting to model a more general 'switch' in the style of speech. As a basic demonstration, a switch along the lines of 'important' to 'unimportant' will be modelled by adjusting basic parameters for the duration of a whole phrase. This technique

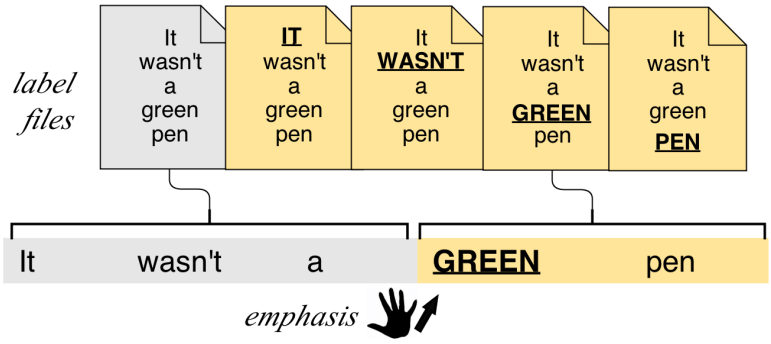
could be extended to changing emotional styles and so on in any future evolution of this work.

3.1.3 Realisation of prosodic control

Given a set of prosodic effects, two primary options exist in terms of implementation:

1. **Recorded database:** Considering firstly contrast, as has been used in [32], it would be possible to use a specially recorded emphatic speech database. At synthesis time, a number of different label files would be generated. One would be 'neutral' (no special intonation), whilst a series of extra label files would be generated, each one marking a different word as 'emphasised'. As a user gesticulates to emphasise a word, the label reader will switch to the appropriate label file in order to emphasise the correct word. Additional databases could be used to train the synthesiser in prosodic effects other than contrastive emphasis. This process is illustrated in Figure 3.1.

Figure 3.1: *Illustrative example for label file switching using pHTS*



2. **Shifts in parameters:** Alternatively, we may control sentence prosody through the live, manual adjustment of parameters such as pitch, duration, energy, and so on. These parameter adjustments may either be drawn up by hand, or learnt from data.

Decision: The first iteration of this project will use manual ***shifts in parameters*** to control prosody. Although both options outlined are feasible, it was felt that manual parameter adjustments would allow a simpler implementation to fit in with the project's time scale. Additionally, modelling prosody in this way is a more interesting task from the author's point of view. It would be interesting to extend the work in future to using recorded speech databases, and comparing the results.

3.1.4 Parameters to alter

A variety of parameters may be altered in order to simulate prosody adjustments. These include **pitch, volume, duration, roundness of vowels, pause model** and **spectral energy**.

Decision: This first iteration of the system will modify ***pitch*** and ***duration*** only. As covered within the literature review, both are very important features in terms of controlling prosodic expression. Furthermore, the previous work upon which this project is built already modifies pitch and duration as part of the demo [8]. Controlling additional pHTS parameters would increase the workload significantly, for a potentially small gain.

3.1.5 Choice of motion sensor as input

A number of suitable motion sensors exist on the market. Two in particular are considered with regards to this project.

1. **Microsoft Kinect:** The Kinect is a motion sensing input device powered by an infra-red projector, camera and 3D scanner system. Since 2011, Microsoft has supported an SDK for developers, more recently C++, C# and Visual Basic code has also been provided as open source[1]. The Kinect is best suited to larger body movements, although next generation sensors are expected to provide significantly better hardware and tracking capabilities, thus may also excel on more subtle gestures[2].

2. **Leap Motion:** Leap Motion is a controller designed to sit on a desktop and track hand and finger movements. The device contains two cameras and three infra-red LEDs. Leap contrasts with Kinect in that the 'observation area' is much smaller, with a much higher resolution (claimed to be 0.01mm)[3].

Decision: The ***Kinect*** has been chosen as the device to be used as the motion sensor for this project. One reason is that the work which this project aims to extend [8] alteady uses the Kinect to control a pHTS voice. Utilising this previous work

[1] http://blogs.msdn.com/b/jgalasyn/archive/2013/03/12/kinect-for-windows-developer-blog-launches-with-announcement-of-new-open-source-sample-code.aspx

[2] http://blogs.msdn.com/b/kinectforwindows/archive/2013/05/23/the-new-generation-kinect-for-windows-sensor-is-coming-next-year.aspx

[3] https://forums.leapmotion.com/showthread.php?1005-Technical-specifications&p=5795&viewfull=1#post5795

should allow more time to be spent on other aspects of the system. Additionally, many resources already exist for motion and gesture recognition within Kinect's Software Development Toolkit, which this project intends to utilise.

Another important factor to consider is whether gestures used to control speech would be more 'natural' to a user if involving large body movements (arm beats, shrugs, etc.) or if involving smaller hand-based gestures (wagging of fingers, turning of palms, etc.). This question has not been addressed further within this project, but the 'naturalness' of particular gestures may be an interesting area for future research. Regardless, it is believed that the larger gestures more suited to the Kinect are not *unsuitable* in themselves, so the Kinect has been chosen based on the previous factors mentioned.

3.1.6 Gesture recognition

There are two primary types of gesture recognition system that may be built using the Kinect, each with their own advantages and disadvantages.

1. **Rule-based:** A rule based system can use Kinect's skeletal tracking functionality to compare the respective x, y and z coordinates of relevant joints. If the relative position of these joints satisfy some condition (either instantaneously or over a period of time), a particular gesture is recognised. The main advantages of such a method are in its simplicity and speed. No advanced algorithms are required, only simple if-then rules. The main disadvantage is its inflexibility. Different users' body shapes may lead to variability in behaviour, and new gestures can't be 'learnt' from users - every gesture must be coded by hand.

2. **Machine learning-based:** Various machine learning methods are suitable in recognising gestures from time series data from the Kinect (SVMs [58] and nearest neighbour classification [57] for example). Features used include angular values between certain joints, frame-by-frame differences, and colour/depth information. The advantage of such methods is in their flexibility. A system such as this could train on individual users to customise gestures, maximising accuracy for each user. Additionally, new gestures may be added by a user, which would be particularly important from an accessibility standpoint.

Decision: For this prototype, ***rule-based methods*** will be used to recognise gestures, for the purposes of simplicity and speed. However, for the reasons mentioned above it

would be desirable to use machine learning methods for more advanced iterations of this work.

3.1.7 Choice of gestures

As discussed within Section 2.5, beat gestures are the most suitable type of gestures to trigger prominence effects in prosody. Within this prototype, large gestures are preferable to smaller gestures in terms of ease-of-recognition. As also previously discussed, it is also desirable to keep gestures as 'natural' as possible, to ensure the device is comfortable and intuitive for the user. The gestures used for each of the planned prosodic effects are now outlined:

1. ***Emphasis* - one handed beat:** One handed beats are hypothesised to work well for emphasis. A neutral position would involve standing with arms by the sides. Considering the apex of the gesture must coincide with the pitch accent, the system will be able to recognise the raising of the arm as anticipating a possible pitch accent. Each arm can be used for a different type of emphasis - a general pitch accent with the right arm, and contrastive emphasis with the left arm.

2. ***Yes/No and WH- questions* - head tilt:** Shoulder shrugging was initially considered to be the most natural 'questioning' gesture. However, testing has shown that the Kinect is currently not sensitive enough to shoulder movements for this to be possible. Therefore tilting the head to one side - sometimes associated with confusion or uncertainty - is used as a next best option. Left and right tilts will refer to different interrogative contours. One handed beats can be used in conjunction to trigger pitch accents within the interrogative contours.

3. ***General importance* - arm width:** Opening and closing the arms in front of the body will be used to indicate general importance (wide arms for higher, slower speech) and unimportance (clasped hands for lower, faster speech). These intend to reflect natural body positions that accompany these types of speech.

4. **Other:** Other gestures considered but not used within this prototype include head nods and shakes, single hand movements to the left and right, wrist flicks, and crossed arms.

3.1.8 Natural language model

As a user gesticulates, the system will recognise a gesture and apply some prosodic effect to the output speech. However, the system may need to make an assumption about exactly which syllable the user intended to apply the adjustment to. For this we can use a natural language model. Various possibilities are outlined here, using the example of emphasis.

1. **No pre-determined bias:** The system would not bias the emphasis toward any particular syllable or word, and would instead place the prosodic adjustment exactly where the user's gesture indicated it to fall.

2. **Stressed syllables:** The label file contains information on which syllables within the synthesis are stressed, and which are not. This information can be used to restrict emphatic peaks (for example, pitch accents) to stressed syllables, as should be the case.

3. **Content vs. function words:** Also contained in the label file is information on whether each word is a 'content word' or 'function word'. This information could be used either to weight probabilities towards content words, or to restrict emphasis to content words entirely.

4. **Use POS tags to bias:** Different parts-of-speech may indicate different probabilities of emphasis being placed upon them. These could be used to bias the system on where emphasis is placed. If a gesture was timed exactly between a noun and an adverb, and adverbs were found to be 1.5 times more likely to be emphasised than nouns, the system would place the emphasis on the adverb.

5. **Use n-grams to bias:** As detailed in [3], unigram/bigram probability of words correlates with accent - the less probable a word, the more likely it is to be accented. Directionally, this concept can likely be extended to emphasis. Thus we could use n-gram probabilities - as analysed by some linguistic engine attached to the synthesiser - to bias the estimated emphasis towards less likely words.

6. **Use a discourse model to bias:** Semantic knowledge - such as identification of whether a word is new to the discourse - could be used to bias emphasis position. However, it has been shown ([59]) that performance using discourse models (of the time) did not improve performance over simpler features such as n-grams, described above.

Decision: For the purposes of this pilot, it has been decided to begin by using information on ***content vs. function words*** (only emphasising content words), and within these, ***information on stress*** (pitch accenting stressed syllables only). There are various reasons for this choice. Firstly, this information is the most readily available in the label files. It will not be trivial to build a more complex language model on top of this. Secondly, the more complex the language model gets in terms of predicting where the emphasis 'should' be, the less need there is for a system that can be controlled by the user (and the harder it would become to emphasise 'unexpected' words).

One issue ultimately found with the method of using content vs. function information is that a small number of 'function' words *do* lend themselves to being emphasised in natural discourse - for example, *'this'* and *'that'*. In cases such as these, an exception list for the system would need to be drawn up by hand, so that the user has the option of emphasising these words. This has not been implemented as part of this pilot.

3.2 Preliminary tests to establish key parameters

Having made preliminary decisions on the key design elements as discussed, a number of parameters must be found - for example, by how much should pitch and duration be modified for each prosodic effect, and how does the natural timing of gestures relate to the prosodic effects? The most rigorous way to establish these parameters would be to carry out regressions based on a large number of recorded phrases. However, given the time available for this project, it has been decided to manually set approximate parameter values based on trial and error. What follows are a number of basic analyses carried out by the author to answer such questions. The small-scale nature of the tests mean that evidence must be taken as directional (and is subject to tweaking once the system has been created), though the data does provide a useful starting point for coding the system's prosodic parameters.

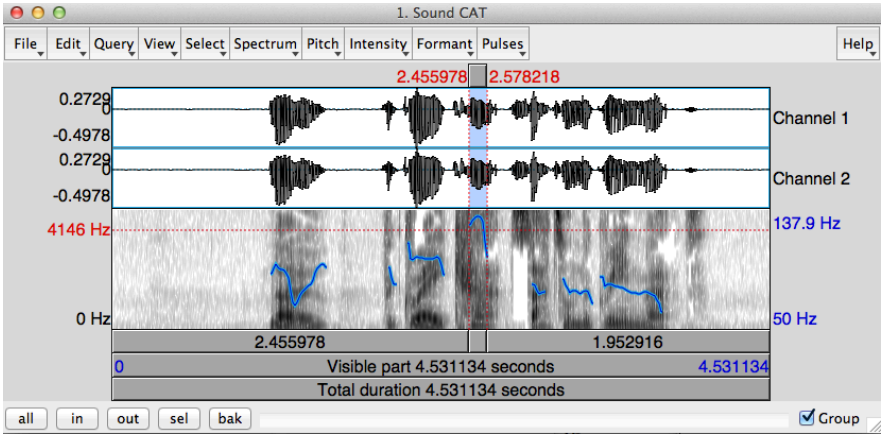
3.2.1 Contrastive emphasis - analysing pitch and duration shift of audio recordings

The author recorded a short sentence of one-syllable words - '*The grey cat sat on the green mat*' in a number of emphatic styles (with a Samson C01 microphone at 44.1kHz stereo using Audacity[4] software). Four neutral versions were recorded, as

[4] http://audacity.sourceforge.net/

well as five contrastively emphasised versions (one emphasising each content word in the sentence). For example, '*The GREY cat sat on the green mat*' was spoken as if in disagreement with '*The white cat sat on the green mat*'. Each audio clip was imported into Praat[5] and each word analysed for pitch (Hz) and duration (sec) with values taken as averages across the totality of the voiced part of each word. These values have been compared to the average values for the four neutral versions.

Figure 3.2: *Using Praat software for pitch and duration analysis.*



Pitch shift results As outlined within Section 2.4, contrastive prosody is often known to take the form H* L-L% or L+H* L-L%. The H* L-L% shape can be seen clearly in the results gathered here. The pitch accent for each emphasised word is very prominent, and is always followed by a decreased pitch relative to the neutral prosody. Figure 3.3 shows the pitch shifts relative to the neutral prosody for the emphasis of each word.

In addition to the clear peaks in pitch and subsequent lowering following the emphasis, there also appears to be a slight raising prior to the emphasis (on average). This information is shown in Table 3.1. These averaged results align reasonably with [28] (whose work has been discussed previously), in which a similar experiment found a contrastive pitch accent to be 28 percentile points higher in value than the same word uttered neutrally.

Finally, it is important to note the fact that the system under design is *reactive*, meaning that alterations in prosody cannot be initiated prior to a gesture being performed. Thus any change in pitch prior the pitch accent is not something the system

[5] http://www.fon.hum.uva.nl/praat/

Figure 3.3: *Pitch shifts (relative to neutral) for 'The grey cat sat on the green mat', spoken with contrastive emphasis on each word in turn. There are clear peaks in pitch on the emphases, general raising in pitch prior to the emphases, and lowering in pitch following the emphases. This information is summarised in Table 3.1.*

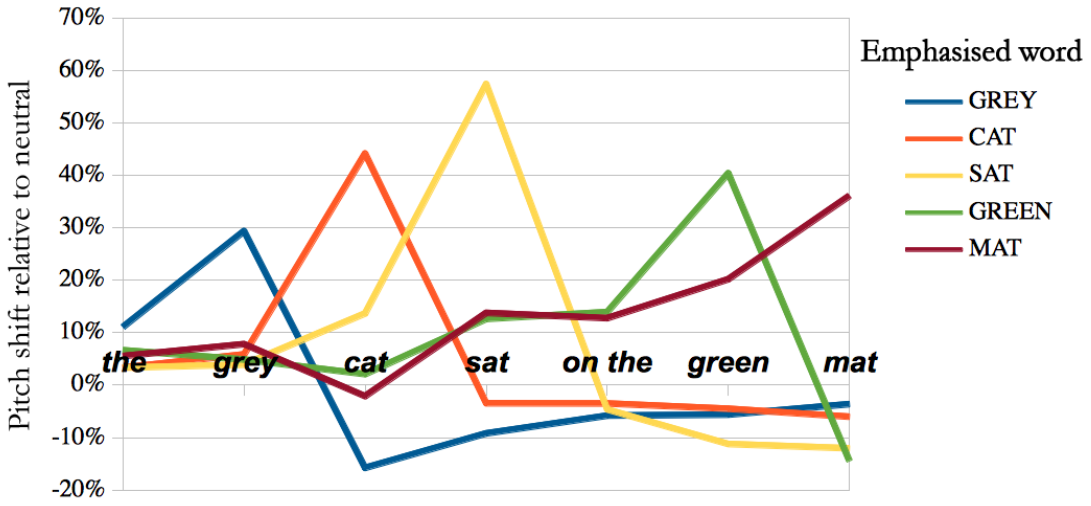


Table 3.1: *Summary of pitch shifts for contrastive emphasis*

	Pre-accent average	**Pitch accent**	**Post-accent average**
Pitch shift:	+8%	+42%	-8%
Standard deviation:	6%	10%	4%

can model (since it cannot anticipate when the emphasis will come). However, both the pitch accent itself, and the subsequent lowered pitch, can be modelled.

Duration shift results As was also noted in Section 2.4, emphasis is also marked with an increase in duration. Tests here have found this to be the case. It can be seen that the emphasised syllable is increased in duration, the preceding word is also generally increased in duration, whilst words prior and after are uttered slightly quicker on average. This data is displayed in Figure 3.4, and summarised in Table 3.2.

The fact that the lengths of syllables are generally shortened in a contrastive sentence intuitively makes sense, as this directs attention away from the pre-supposed information and towards the focus of the sentence. The decrease in speed (increase in duration) prior the emphasised word may function as preparing the listener for an important word.

Figure 3.4: *Duration shifts (relative to neutral) for 'The grey cat sat on the green mat', spoken with contrastive emphasis on each word in turn.*

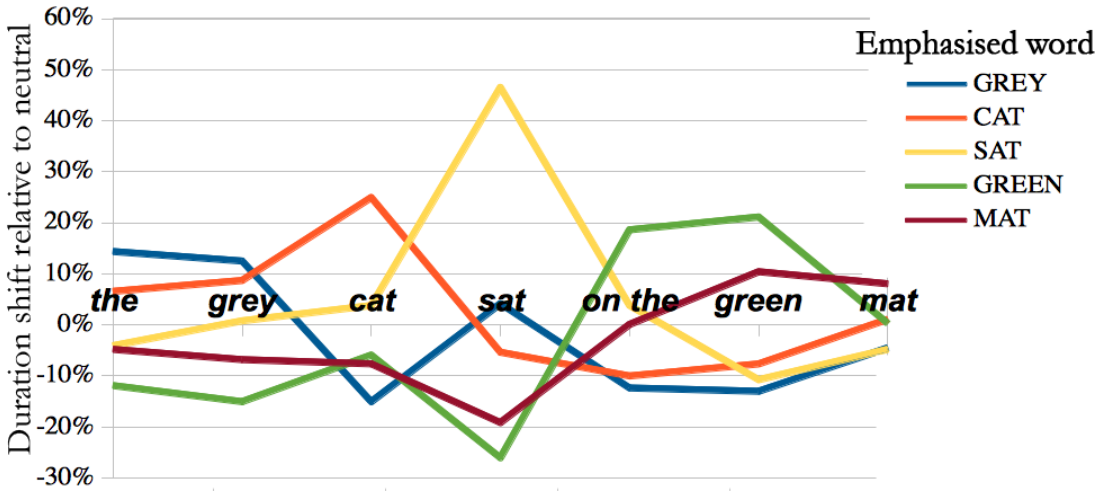


Table 3.2: *Summary of duration shifts for contrastive emphasis*

	Pre-emphasis (excl prev word)	**Preceding word**	**Emphasised word**	**Post-emphasis**
Duration shift:	-8%	+11%	+23%	-6%
Standard deviation:	9%	6%	14%	6%

3.2.2 Reverse engineering contrastive emphasis

Having explored the pitch and duration effects of contrastive emphasis as just described, a complementary way to continue testing is to manually adjust the intonation of neutral synthesis in order to recreate a 'natural' sounding contrastive emphasis effect.

A standard male Scottish voice from Mage v2.0[6] has been used to synthesise the sentence *'She found herself falling down a very deep well'* (from Lewis Carroll's Alice in Wonderland). This sentence was chosen as the label file already existed within the download package. The resulting sound file has been manipulated multiple times in Praat, with the pitch adjusted to simulate contrastive emphasis as best possible on each of the content words.

Results generally align with previous findings: the emphatic peak should be raised (30% appears to be a more reasonable on the synthesised text, rather than 40% or

[6] http://mage.numediart.org/

more), and the subsequent remainder of the sentence lowered in pitch by 20%. In some cases naturalness was increased by increasing the pitch of words prior the the pitch accent. Additionally, the final word of the sentence (*'well'*), when neutrally synthesised, falls in pitch throughout (pHTS synthesises it with a slight emphasis). In order to simulate realistic contrastive emphasis on a preceding word, this fall in pitch needed to be 'flattened out' manually. These results are detailed in Table 3.3.

Table 3.3: *Pitch shifts manually added to synthesised text to simulate contrastive emphasis. * indicates the pitch was 'flattened' out to remove the small default pitch accent.*

	FOUND	**HER-**	**FALL-**	**DOWN**	**VER-**	**DEEP**	**WELL**
she	0%	0%	10%	0%	0%	0%	0%
found	**30%**	10%	10%	0%	0%	0%	0%
her-	-20%	**30%**	10%	0%	0%	0%	0%
-self	-20%	-20%	10%	0%	0%	0%	0%
fall-	-20%	-20%	**30%**	0%	0%	0%	0%
-ing	-20%	-20%	-20%	10%	0%	0%	0%
down	-20%	-20%	-20%	**30%**	10%	0%	0%
a	-20%	-20%	-20%	-20%	20%	0%	0%
ver-	-20%	-20%	-20%	-20%	**30%**	0%	0%
-ry	-20%	-20%	-20%	-20%	-20%	10%	0%
deep	-20%	-20%	-20%	-20%	-20%	**30%**	10%
well	-20%*	-20%*	-20%*	-20%*	-20%*	-20%*	**40%**

3.2.3 Other prosodic effects

The methods demonstrated so far provide directional guidance on how parameters should be shifted to simulate *contrastive emphasis*. Although informative, the process is time-consuming, and to maintain a reasonable project scope, parameters for other prosodic effects (general emphasis, interrogative contours, etc.) have been set by ear, without analysing or manually manipulating sound files using Praat. However, the essence of the methodology is the same, in that basic intonation shapes for prosodic effects are noticed, and parameters are tested on speech synthesised by the system. Parameters are then tweaked within the code to result in the most natural sound possible.

3.2.4 Beat gesture timing

An analysis to observe the timing of a natural beat gesture - similar to that performed in [56] - has been carried out by the author. The phrase (*'She found herself falling down a very deep well'*) was spoken seven times, with each content word emphasised in turn (both vocally and with a large beat gesture). Audio and visuals were recorded using a Canon EOS D500, and played back frame by frame using Apple's iMovie.

Five points in the arm motion were tracked with respect to the speech: the start of the motion, the point at which the wrist crosses the elbow vertically, the point at which the wrist is level with the shoulder, the peak of the motion, and the bottom of the downward beat (illustrated in Figure 3.5). These are aligned with a syllable in the speech ('0' representing alignment with the pitch accent, '-1' representing alignment with the syllable prior, etc.). Averages from all attempts were taken, and the temporal relation between gesture and pitch accent can be seen in Table 3.4.

Figure 3.5: *Tracked arm motion: from left to right,* ***start of movement, hand crosses elbow, hand crosses shoulder, top of peak*** *and* ***bottom of beat****.*

Table 3.4: *Number of syllables by which beat gesture is offset from pitch accent (negative numbers indicate movement prior to pitch accent)*

	Mean (# syllables)	**Median (# syllables)**
Start of movement (hip)	-2.0	-2
Hand crosses elbow	-1.4	-1
Hand crosses shoulder	-0.9	-1
Hand at peak of movement	-0.3	0
Hand at bottom of beat	+0.3	0

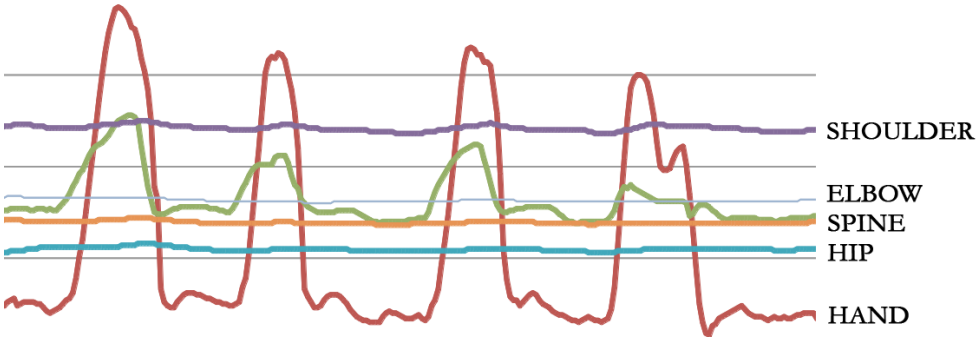
Results show that the arm movement begins (on average) 2 syllables prior to the pitch accent. Thus the latency of the system cannot be longer than this, unless we wish the user to perform a slower motion, starting the movement earlier. The peak of the beat falls on average 0.3 syllables prior to the emphasis (i.e. either on the emphasis, or the syllable prior).

3.2.5 Beat gesture recognition

Having established that gestures will be recognised through simple rules (rather than time-based machine learning methods) as described in Section 3.1.6, algorithms to trigger emphasis based on relative joint positions must be written.

The most significant gesture to be used by this system is the 'beat' gesture previously discussed, and it is this gesture that is focused on here. In order to work out a simple and effective way of recognising such a gesture, the author repeated a beat gesture in front of the Kinect, whilst tracking the x and y coordinates of the hand, elbow, shoulder, hip and lower spine. Whilst the x coordinate contains little useful information in this case, the y coordinate can be plotted over time, as shown in Figure 3.6.

Figure 3.6: *Vertical (y) coordinates of various joints as tracked by the Kinect throughout four beat gestures. The red and blue coordinates cross as the hand moves vertically past the hip.*



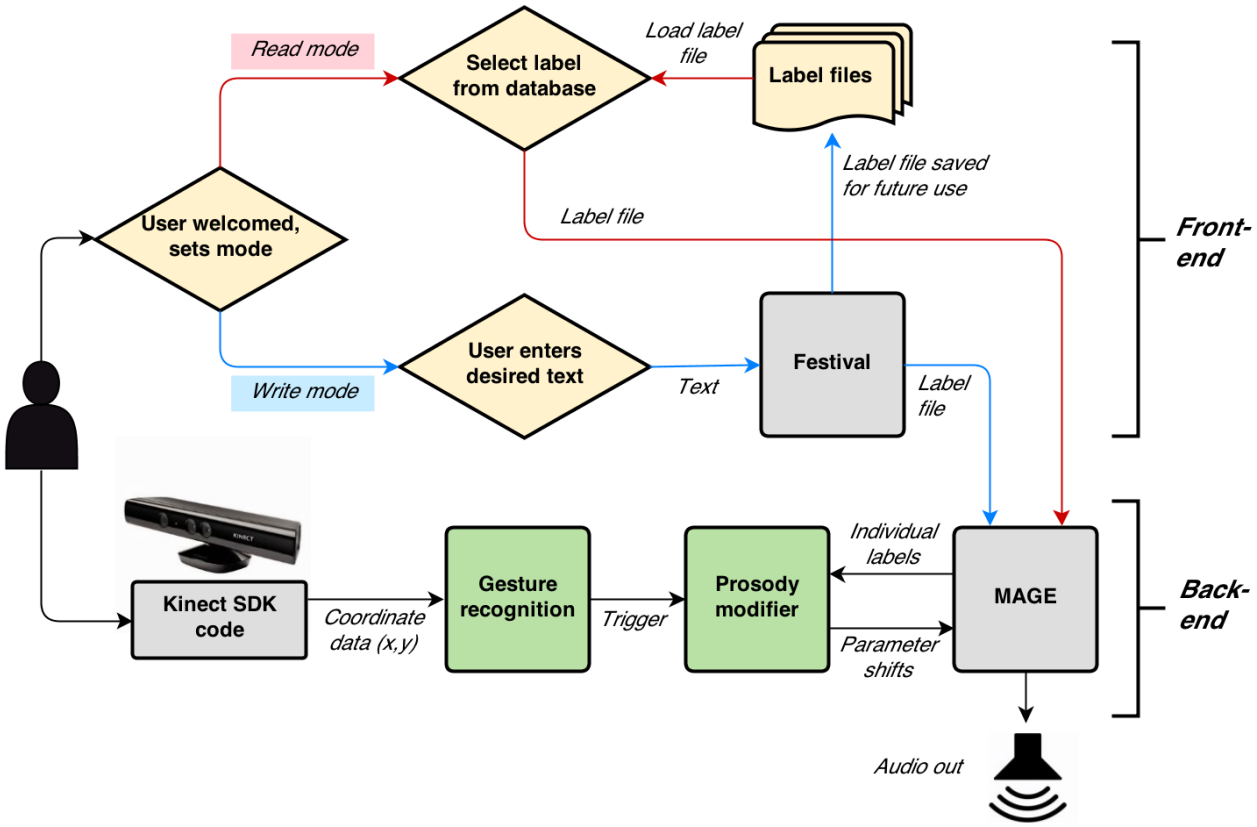
In terms of gesture recognition, the goal of our system is to recognise with reasonable certainty that the user is performing a gesture *as early as possible*. From the figure it can be seen that one of the earliest signals that a beat gesture is being performed is the vertical crossing of the hand (red) with the hip joint (light blue). Although we cannot be sure that this condition will exclusively distinguish a *beat* gesture once further gestures are added to the system's capabilities, for the purposes of this pilot, the vertical crossing of the hand and hip can act as a simple trigger that the user intends to

emphasise an upcoming word.

3.3 Implementation

The following schematic (Figure 3.7) illustrates the basics of how the system is constructed. A frontend, created in Python, allows the user to enter or load text to synthesise, which is sent the the backend as a label file. The backend, written in C++, synthesises this text using the pHTS engine, whilst tracking the user's skeleton. As a gesture is recognised, the pitch and speed parameters used within the pHTS engine are modified according to a set of well defined rules. The following sections provide further detail on the system setup, and the specific gestural/prosodic rules implemented.

Figure 3.7: Schematic diagram representing system setup. Grey shading represents pre-existing components. Yellow represent elements created for the frontend, green represent elements created for the backend.



3.3.1 Frontend

The system has two modes, **write mode** and **read mode**. The purpose of both modes is to create a temporary label file (based on some user input) which the backend loads into MAGE. The format of the label file required by the system is that outlined within [60]. Features contained include the current and surrounding phonemes, stress on syllables, position of syllables and stressed syllables within the phrase, whether words are content or function words, and so on.

In **write mode**, the user's textual input is fed to **Festival** Speech Synthesis System v2.1[7] using Pythons `subprocess` module. Parameters are passed to Festival to create the appropriate label file (using *voice_cmu_us_slt_arctic_hts*) and to use a CART tree to predict phrase breaks[8]. The temporary label file created by Festival is extracted from the system's *Temp* directory, the format is cleaned, and it is saved to a set location in order to be loaded by the backend. In **read mode**, a label file already created previously by this process is selected by the user, and copied to the same set location to be loaded by the backend. In both modes, two additional files are also created - one containing the sentence text in plain English, and another containing the length of the label file (both are required by the backend).

3.3.2 Backend

The backend is built in C++ on top of MAGE and the work already carried out as part of [8]. OpenFrameworks[9] is used as the graphical and audio framework.

The main application loop repeatedly calls an `update` function (15 times per second), in which the Kinect 'skeleton' is processed by pre-existing **Kinect SDK code**. This sets variable parameters for *x* and *y* coordinate data for all tracked joints.

The updated coordinate data for joint variables are used within the **gesture recognition** stage. A simple set of `if-else` rules act as triggers. These are described in more detail in Section 3.3.4.

Once some gesture has been recognised, the **prosody modifier** must allocate the prosodic effect to a specific syllabic unit (or set of syllabic units). Again, this is carried out using a set of `if-else` rules drawn up by hand, a set of magnitudes for parameter shifts, and information from the label file (again, see Section 3.3.4 for specifics). The current label is fed back via a pointer from MAGE, and parsed using regex.

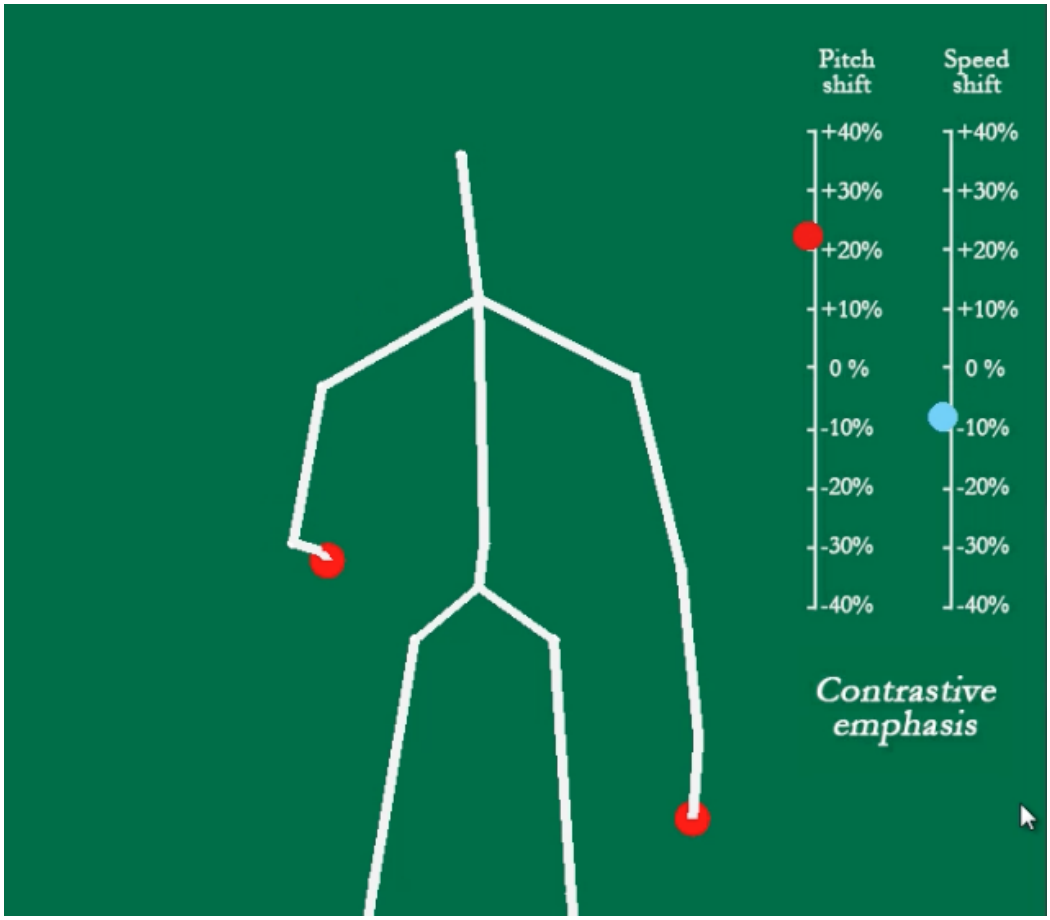
[7] http://www.cstr.ed.ac.uk/projects/festival/download.html
[8] http://www.cstr.ed.ac.uk/projects/festival/manual/festival_17.html
[9] http://www.openframeworks.cc/

The parameter shifts are sent to the **MAGE** engine via the functions `MAGE_setPitch` and `MAGE_setSpeed`. These are used to shift parameter trajectories as appropriate via the pHTS engine. This functionality pre-exists within the MAGE platform code.

3.3.3 Additional visual output

In addition to the audio out, the application provides a visual representation of the user's 'skeleton' as part of the application interface. Flashing text indicates to the user in real time if a prosodic effect has been triggered, and two graphical meters representing pitch and speed shifts indicate the parameter shifts being applied at any moment. Note that the majority of this functionality has been added as an extension in the final week of the project, and only the simple skeletal feedback was functional within the testing phase. A screenshot can be seen in Figure 3.8, and live output viewed in the video demonstration at *http://tinyurl.com/msc-synthesis*.

Figure 3.8: *Screenshot of the final system in action*

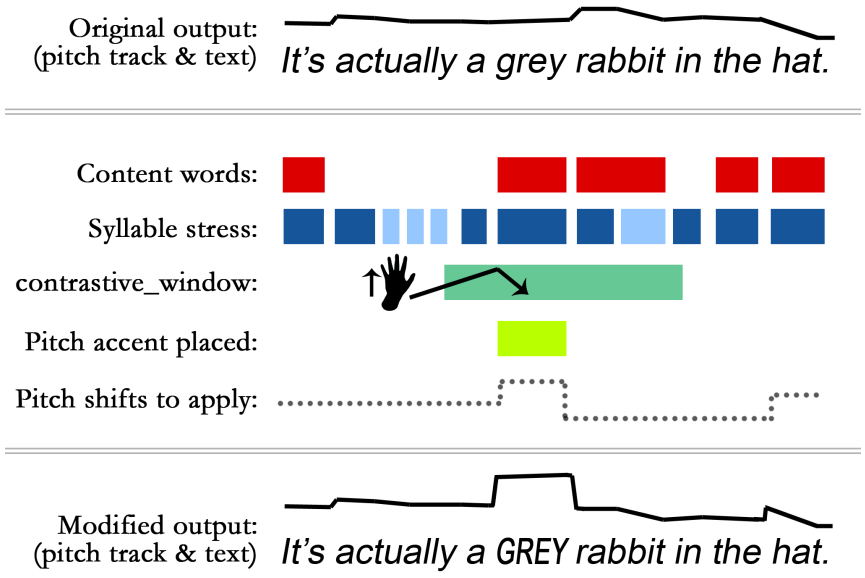


3.3.4 Prosodic rules implemented

This section details the specific rules used within the code to implement the prosodic effects present in the system, along with illustrative examples. Pseudocode, in the form of flow-charts, are presented in the Appendix. All parameter values have been set using methodology discussed previously in Section 3.2 and fine-tuned by ear.

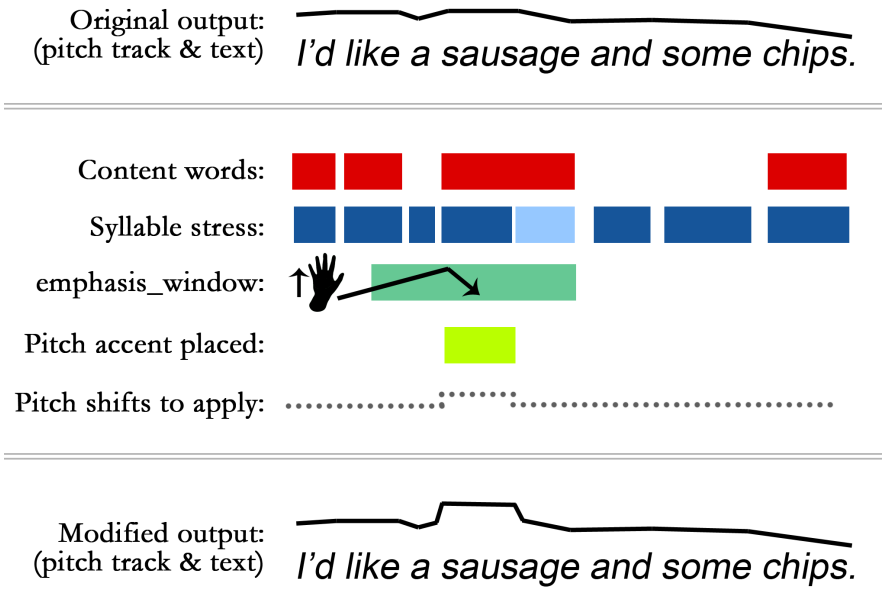
Contrastive emphasis

- **Gesture rules:** Due to latency issues that haven't been possible to resolve (discussed further in Section 3.3.5), contrastive emphasis is triggered as the left hand moves above the left hip, but no coordinate data for the hand above this point can be used to fine-tune the pitch accent position. A window ('*contrastive_window*') of 8 frames (∼0.5 seconds) is triggered as the left hand passes the hip, in which a pitch accent may be applied should the prosodic rules allow.
- **Prosody rules:** A pitch accent is applied if a content word's stressed syllable falls within this 0.5 second window. The pitch accent consists of a raised pitch (28%) and a reduction in speed (-10%). Following this, the remainder of the sentence is lowered in pitch (-14%) and increased in speed (14%). The final syllable of the sentence is raised in pitch (10%) to counter the pHTS default falling accent.
- **Flowchart:** See Appendix A.
- **Illustrative example:**



General emphasis

- **Gesture rules:** As with contrastive emphasis, a 0.5 second window ('*emphasis_window*') for general emphasis is triggered with a hand movement above the hip (though this time, the right hand and hip are used).
- **Prosody rules:** A pitch accent is applied if a content word's stressed syllable falls within this window. The pitch accent consists of a raised pitch (15%) and a reduction in speed (-10%). Following this pitch accent, the rest of the sentence is unaffected.
- **Flowchart:** See Appendix A.
- **Illustrative example:**

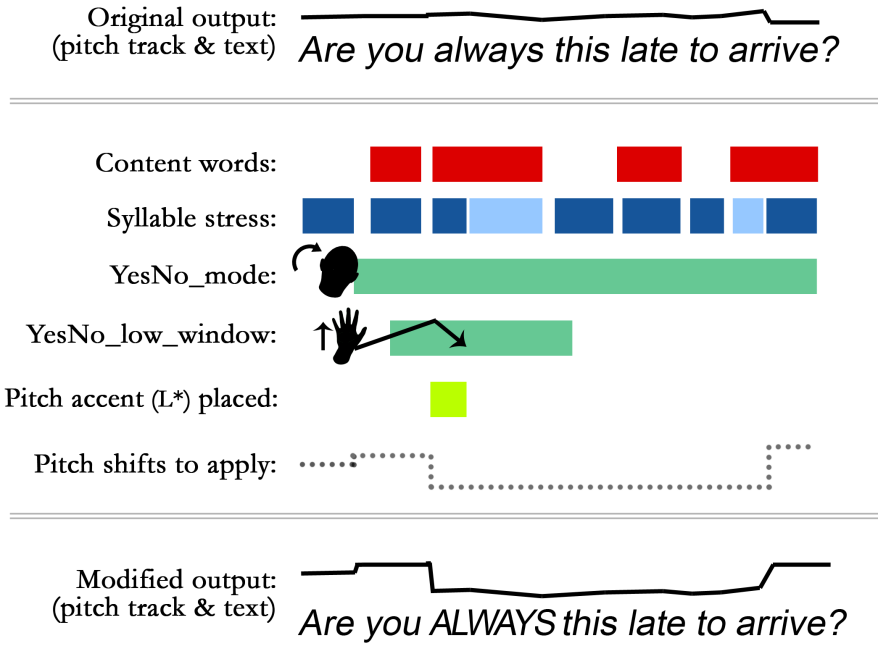


Yes/No question

- **Gesture rules:** A head tilt to the right triggers prosodic intonation for a Yes/No question. This movement should be carried out close to the start of the sentence. At some point in the sentence, the user can emphasise with their arm (as described above, bringing the arm above the hip). The left arm results in an emphasised word with a high (H*) accent, the right arm with a low (L*) accent.
- **Prosody rules:** The head tilt applies an immediate increase in pitch (10%), and guarantees that the final syllable will rise in pitch (25%) for the required high

boundary tone (H%). If the user performs a beat gesture with their left arm, a high pitch accent is applied to the next content word's stressed syllable (20%), before the pitch is lowered for the phrase boundary prior to the final syllable (-15%). Alternatively a beat gesture with the right arm (instead of the left) leads to a low pitch accent (-15%) followed by the low phrase boundary (-15%).

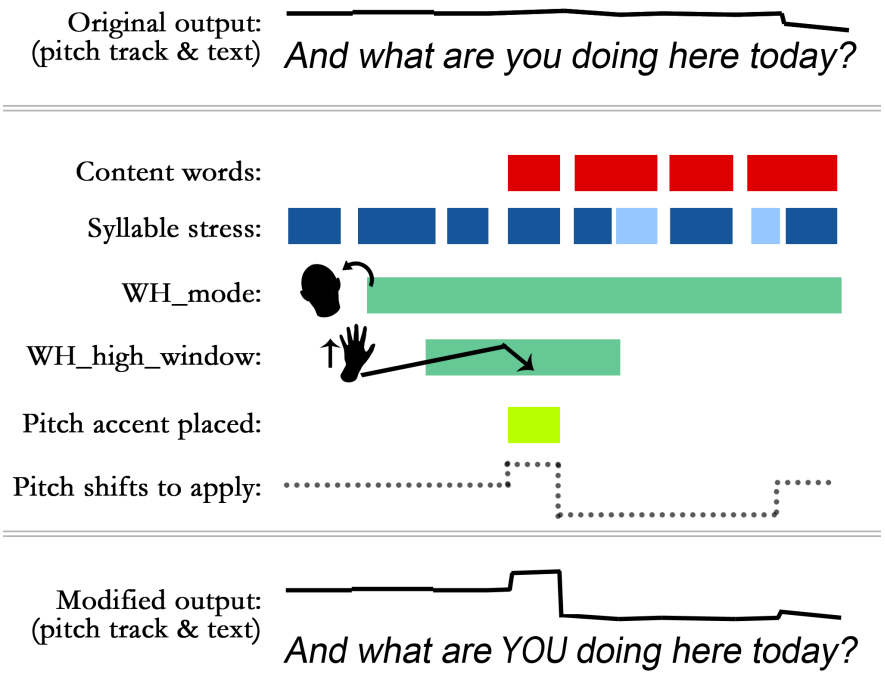
- **Flowchart:** See Appendix A.
- **Illustrative example (using a low pitch accent):**



WH-question

- **Gesture rules:** A head tilt to the left triggers prosodic intonation for a WH-question. This movement should be carried out close to the start of the sentence. At some point in the sentence, the user must emphasise with their left arm (as described above, bringing the arm above the hip).
- **Prosody rules:** The head tilt applies an immediate increase in pitch (10%). When the user performs a beat gesture with their left arm, a high pitch accent (H*) is applied to the next content word's stressed syllable (35%). The remainder of the sentence is lowered in pitch (-12%) representing the low phrase boundary and boundary tone (L-L%). As with contrastive emphasis, the final syllable is raised in pitch (10%) to counter the default falling pitch accent that pHTS normally applies to the final syllable.

- **Flowchart:** See Appendix A.
- **Illustrative example:**



3.3.4.1 'Importance' and 'unimportance'

- **Gesture rules:** As the horizontal distance between the user's hands is decreased to less than shoulder width, the 'unimportant' mode is triggered. As the horizontal distance between the user's hands is increased to greater than 2.5 times the shoulder width, the 'important' mode is triggered. Both modes are cancelled as soon as the distance between the hands no longer crosses either threshold.
- **Prosody rules:** 'Unimportant' mode results in a decrease in pitch (-10%) and an increase in speed (+10%). 'Important' mode results in an increase in pitch (+10%) and a decrease in speed (-5%).

3.3.5 Implementation issues and discussion

MAGE version Although it was attempted to use MAGE 2.0 to build the system, unresolved issues meant that code based upon MAGE 1.00 was ultimately used. Although MAGE 1.00 acts as a prototype, allowing control over parameters as seen in this project, MAGE 2.0 claims to be more reactive, with every phoneme processed individually (rather than on a sliding label of two phonemes). Additionally, an audio thread

processes each sample individually, whereas MAGE 1.00 buffers groups of samples. Thus the control of MAGE 2.0 is said to be *'much more reactive and accurate'* [61]. Future versions of this work should therefore attempt to use the MAGE 2.0 framework, which will result in a system that can be controlled to a more accurate level.

Latency and its implications As previously noted, one of the main issues encountered has been in the relatively poor latency of the system. It would have been desirable to use the whole trajectory of the rising arm to alter prosody, 'honing' in on a syllable to emphasise as the user's hand reaches its vertical peak. As it is, the system must make do with a hand passing the hip as a signal that emphasis is likely to occur soon, which puts much more onus on the user to move their hand in a very specific way, so that the initial raise of the hand past the hip takes place at the correct time to emphasise the desired word.

It appears that this issue is a result of the audio buffering that MAGE 1.00 utilises. The Kinect processing carried out in C++ can be seen to occur in near real-time (through the on-screen skeletal tracking), and MAGE itself applies parameter changes with a delay of just one label. However, even in the simple case of mapping a hand's y coordinate to pitch, there is a multisyllabic delay between hand movements and the audio being altered, which suggests the issue is one of audio buffering. Thus from the user's perspective, although they may make a gesture which would seem to be simultaneous with a word emitted from the system's speakers, the system itself is currently processing multiple syllables ahead of this in the sentence. Thus the start of the user's movement must be performed suitably ahead of the desired emphasis.

Another issue resulting from this large latency is the fact that the prosodic effects modelled need to be simplified slightly relative to what they could be with smaller latency. For example, the H* pitch accent within contrastive emphasis is often described as L+H* (as outlined previously). If the latency of the system were improved, there would be more scope to alter prosody prior to the emphatic pitch accent - for example, lowering the preceding syllable. Likewise, we may want to decrease the speed of speech slightly before the emphatic syllable. However, neither have been implemented, as the hand crossing the hip often only allows just enough time to apply the pitch accent itself, with no scope for altering the syllable prior.

Type of pitch adjustments Although increasing the pitch of a syllable appears to work as a decent approximation to a natural pitch accent, there are some cases in

which the resulting speech does not sound natural. For example, emphasis on the final syllable of a phrase in *natural* speech should not just result in an increased pitch, rather an increased pitch and exaggerated fall. However, this type of effect has not been possible to implement using MAGE. Although the pitch of labels may be shifted linearly, a pitch contour *within* a label cannot be drawn. If this capability were to be added in future, more subtle and realistic pitch contours may be added, such as a falling or rising pitch contour within a single label.

Gesture recognition capability As mentioned within Section 3.1.7, although it was hoped to implement a 'shrug' as the gesture to trigger the interrogative prosody modes, the Kinect's skeletal tracking abilities are not sensitive enough to shoulder movements to make this a possibility. A 'head-tilt' is therefore used as a substitute. Future versions of the Kinect may be more sensitive to shoulder movement, however.

Overall speed It has been found that by setting a baseline speed at 85% of what is normally classed as a 'neutral' speed for pHTS, the synthesis is easier to control. Any slower than this begins to sound increasingly unnatural. Thus all speed shifts referred to in this project are relative to the base speed, running at 85% of the default pHTS speed.

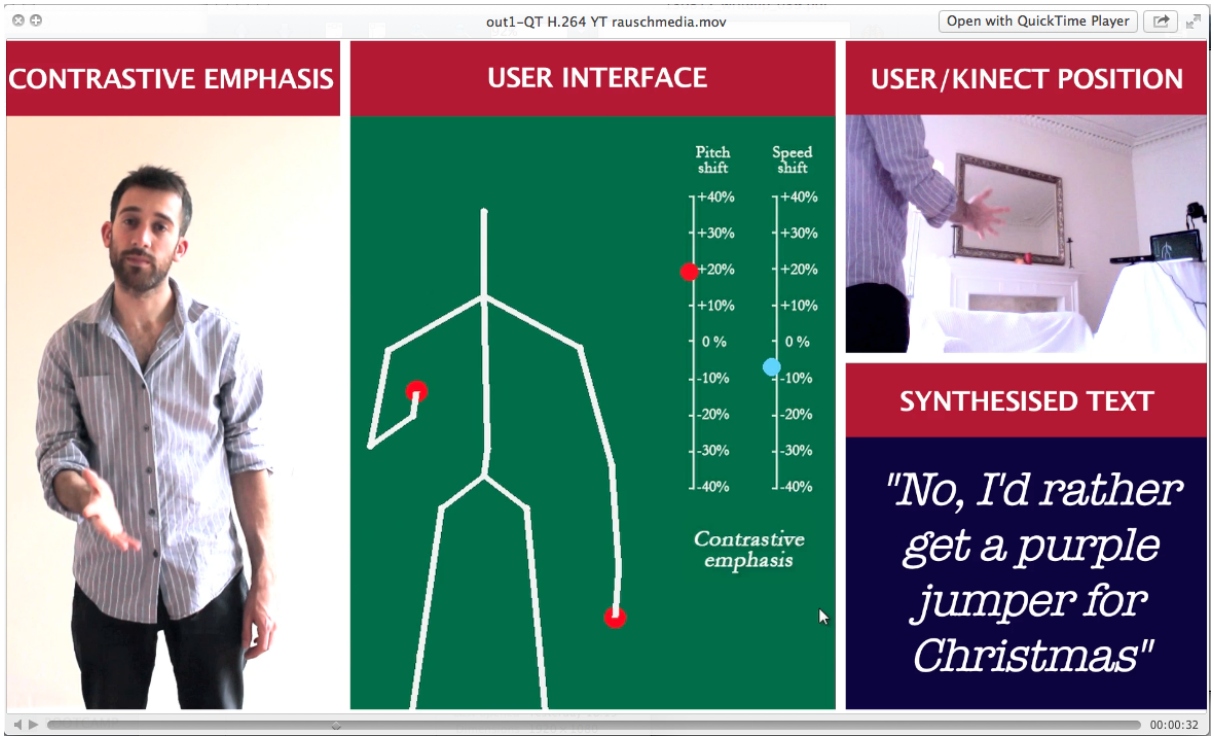
'De-stressing' words As described, the pitch is lowered (and the speed increased) after a contrastive emphasis pitch accent. However, it was hypothesised that some syllables may need to be additionally 'unaccented', in the case that pHTS were to stress these in its neutral form. However, the level of stress added by pHTS to words is minimal, and additional adjustments such as this have not been required. The only exception is in the case of the final syllable of the sentence, which is raised slightly to counter the falling accent pHTS applies naturally.

End of label file glitches The system is currently set to repeat the label file indefinitely, until the user exits the application. However, jumping from the end of the label file and back to the beginning causes the audio to 'click', most likely due to the two 'silent' waveforms not lining up exactly.

3.3.6 Demonstration video

The reader is directed to a demonstration video, which shows examples of the system in action. This provides a clear picture of the gestures that need to be performed, and the specific pitch and speed contours that each prosodic effect creates. This video can be found at *http://tinyurl.com/msc-synthesis*.

Figure 3.9: *Screenshot of the accompanying video demonstration*



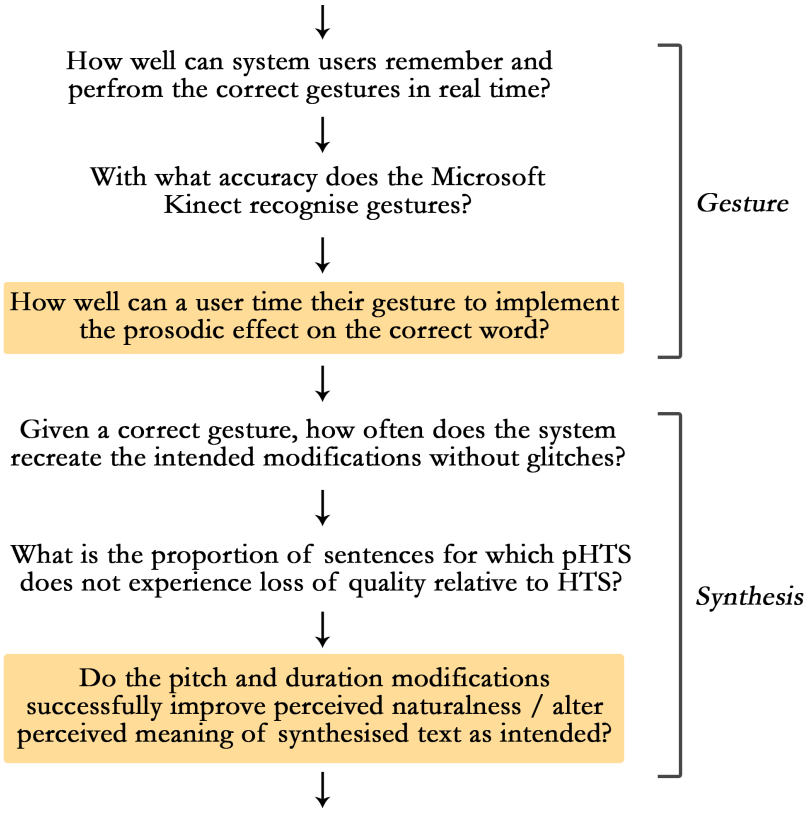
Chapter 4

Areas to test

Chapter summary: *This chapter splits the problem statement into a series of sub-questions. Two in particular are focused on, with testable hypotheses drawn up.*

The original problem statement asks the question of whether it is possible to **generate superior prosody in a speech synthesis system, using natural gestural controls, in real time**. This question is now broken down into a series of sub-questions whose hypotheses may be explored through individual experiments, illustrated in Figure 4.1. This project aims to tackle only the bigger and more fundamental questions through formal tests (shaded in figure).

Figure 4.1: *Flow of sub-questions. This project focuses formally on the shaded items.*



How well can system users remember and perform the correct gestures in real time?

This question is not tested within this project, but would be an interesting area to investigate in future. As the number of prosodic effects available to the user increases, it is hypothesised that the user would take longer to learn the various gestures, and their accuracy rate may decrease.

With what accuracy does the Microsoft Kinect recognise gestures?

Once a user performs a gesture, the Microsoft Kinect must 'recognise' this gesture in order to alter the synthesised prosody. One would expect the 'rule-based' approach used within this project to result in a high accuracy recognition rate. The number of gestures not recognised have been counted across multiple 'generation tests' (outlined subsequently in Section 6.1.2), which provides a directional picture on the precision rate for gesture recognition.

For a more thorough and robust test however, future work should measure both

precision and recall across different types of gestures and under different conditions. For example, different lighting conditions, camera positions, body shapes and types of gestures all affect the ability of the Kinect to correctly identify gestures.

How well can a user of the system time their gesture to implement the prosodic effect on the correct word?

As previously covered, a number of rules are in place to assist the user in adjusting the prosody of specific syllables (i.e. content words only, stressed syllables only). However, it is still possible to emphasise the wrong word if a movement is not performed within some required window of time.

User tests are carried out in which users are required to emphasise certain words with certain gestures. Both quantitative and qualitative results are recorded. The primary metric tracked in the quantitative tests is accuracy of timing. These tests are hereby referred to as **generation tests**.

- **Overall accuracy rate over extended period of time**
  - *Hypothesis: users will emphasise correct words with a greater accuracy than that determined by 'chance'*
- **Change of accuracy rate over course of session**
  - *Hypothesis: users will increase their accuracy rates over the course of a session*
- **Repetition of a sentence**
  - *Hypothesis: users will increase their accuracy rate for a particular sentence with repetition*
- **Position of emphasised word within phrase**
  - *Hypothesis: the position of the word in the sentence will affect a user's accuracy rate (in unpredictable ways)*
- **Naturalness of word to emphasise**
  - *Hypothesis: accuracy rate will be higher for words that are more 'natural' to emphasise than words that wouldn't typically be emphasised in natural speech*
- **Speaking alongside the synthesiser**

  - *Hypothesis: accuracy rate will increase if a user speaks 'alongside' the synthesiser (due to increased awareness of position in sentence and naturalness of gesture)*

- **Number of gestures per sentence and their proximity**
  - *Hypothesis: accuracy rate will decrease if two gestures are performed in close proximity*

Given a correctly timed gesture, how often does the system recreate the intended pitch and duration modifications - i.e. without glitches?

It is the case that the system output can sometimes 'glitch' unexpectedly when trying to alter the pitch on certain words. However, for the purposes of the generation and listening tests carried out, sentences have been filtered so that only those in which *it is possible to emphasise the word in question successfully* are chosen. This is justified by the fact that we are trying to test the users' abilities to emphasise certain words with gestures (in the generation tests), and the quality of the method of modifying pitches and durations (in the listening test). We do not wish to introduce an extra factor of 'words that the system can't deal with' at either of these stages.

A possible extension to this project would be to assess the quality of the system in terms of these glitches. A rigorous test would involve creating lists of sentences to synthesise with each possible gesture, ensuring a balanced set of diphones and prosodic effects are present (similar to the recording process for unit selection synthesis).

What is the proportion of sentences for which pHTS does not experience loss of perceived quality relative to standard HTS?

As described, the system uses pHTS, rather than a standard HTS engine, to synthesise speech. At times, the pHTS output compares adequately with HTS, whereas at other times the quality is noticeably inferior. For example, the *'t'* phoneme is in some situations very muffled. As previously mentioned in the literature review (Section 2.2), the creators of pHTS claim the quality decrease from standard HTS to be 'relatively minor', though this assertion is itself subjective.

In order to attain whether the overall system created increases the overall 'quality' relative to current generation HMM systems, it would be important to assess by how much the use of pHTS over HTS affected the quality, as part of the overall 'pipeline' of sub-questions. However, to maintain a manageable scope, this project will use the

pHTS engine's neutral synthesis as its baseline.

Do the pitch and duration modifications successfully improve perceived naturalness and/or alter perceived meaning of synthesised text as intended?

Assuming an utterance is synthesised as intended, it should be assessed how the prosodic effects are perceived by listeners. To maintain a manageable project scope, listening tests are restricted to tests on two of the system's prosodic capabilities - **contrastive emphasis**, and **low accented interrogative prosody** (i.e. yes/no questions with L* accent). Contrastive emphasis has been chosen as a commonly occurring prosodic effect which should be relatively easy for the listener to notice. Interrogative prosody with a low accent has been chosen as - although rather uncommon - the effect allows us to test how well listeners respond to a much more subtle emphatic effect. High accented interrogative prosody would likely to lead to similar results to contrastive emphasis, which would not be as interesting for the purposes of this project.

The sentences used in testing are all synthesised by the system itself, with the author controlling prosody through gestures. It has been ensured that where appropriate the above factors (i.e. correct gesture, correctly timed, correctly synthesised by pHTS etc.) are all fulfilled. This way results will reflect listeners' opinions on the prosodic contours rather than external factors. However, some parts of the experiment purposely test 'slipped' and mistimed gestures, to assess the impact these have on naturalness from a listener's point of view.

The following briefly outlines the areas to be tested through a listening experiment. Details of test formats are included within Section 5.2.

- **Contrastive emphasis - naturalness:**
  - Emphasis on **correct** word in contrastive sentence/dialogue *(hypothesis: correct emphasis improves perceived naturalness vs. neutral synthesis)*
  - Emphasis on **incorrect** word in contrastive sentence/dialogue *(hypothesis: incorrect emphasis decreases perceived naturalness vs. neutral synthesis)*
  - Emphasis on correct word in contrastive sentence/dialogue though gesture slightly **mistimed** leading to small glitch on emphasised syllable *(hypothesis: slipped emphasis decreases perceived naturalness vs. neutral synthesis)*
- **Contrastive emphasis - semantics:**

  - Emphasis on various words in sentence *(hypothesis: position of emphasis can alter the perceived semantics of a sentence in an intended way, relative to neutral synthesis)*

- **Contrastive emphasis - position in sentence:**
  - Emphasis at various positions in sentence, including one-syllable and two-syllable words *(hypothesis: correct emphasis on the final syllable of phrase decreases naturalness relative to neutral synthesis, emphasis elsewhere in sentence to increase naturalness vs. neutral synthesis)*
- **Low accent interrogative - naturalness:**
  - Low accent interrogative prosody on appropriate question text *(hypothesis: interrogative synthesis with a low accent increases perceived naturalness vs. neutral synthesis)*
- **Low accent interrogative - semantics:**
  - Low accent emphasis on various words in interrogative sentence *(hypothesis: position of low accent within interrogative prosody alters the perceived semantics of a sentence in an intended way, relative to neutral synthesis)*
- **Low accent interrogative - question vs. statement semantics:**
  - Low accent interrogative prosody on statements (which could feasibly be questions) and WH-questions *(hypothesis: interrogative vs. neutral prosody affects the interpretation of declarative text (i.e. whether a listener perceives the speech as a statement or question), but a WH-word within the synthesised text overrides any interrogative vs. neutral prosody)*

If so, what are the optimal magnitudes of pitch and duration shifts in improving perceived naturalness and/or meaning?

Although not strictly part of the sub-question 'pipeline' asking if such a system works, it is important to optimise parameters in order to build the best system possible.

The parameters for the pitch and duration modifications for each prosodic effect have been set through preliminary tests as outlined in Section 3.3.4. However, in order to optimise values it would be desirable to test the effect of tweaking these parameters on subjective tests such as those outlined above. This has not been carried out in depth, though as a demonstration, a small section of the listening test is used to compare perception of naturalness for different values of F_0 in contrastive pitch accents.

Chapter 5

Experimental setup

Chapter summary: *This chapter outlines the experimental setup and conditions for both the generation and listening tests. For clarity, details on specific sentences used in each test are described within the following chapter, alongside the results.*

5.1 Generation test

5.1.1 Setup summary

In total, 12 native English speakers are tested using the system, nine being friends of the author participating for no reward, and the three others receiving £7 in compensation. Each user was asked to add some form of emphasis to 31 different sentences through gestural control, with each sentence being repeated eight times consecutively. This repetition is designed to produce results that see past the 'learning effect' for each sentence, as well as increasing data for the purposes of significance testing. Each sentence contains one or two gestures, leading to a total of 264 gestures that have been performed and tracked for each user.

Each user was provided with a brief introduction as to how the system worked (though no technical details on how gestures are recognised or should be timed), and was stood in front of the Kinect, with the author sat to the side and within view. After the first five sentences (each with eight repetitions), the user was shown in more detail how the Kinect recognises the emphasis gesture, and provided with any further advice on how to increase accuracy. This is hereby referred to in this report as 'training'.

A script was placed within the user's view. For each sentence the user was told which word to emphasise, and with what action (normally contrastive emphasis, as

this is the most obvious to the ear). The author was able to track each attempt as being correct, early / late (without emphasising the wrong word) or very early / very late (emphasising the wrong word). Immediately after each attempt and prior to the next, the subject was told by the author if they had gestured correctly, early or late (although it was often clear to the user without prompting). This feedback is justified, as in a fully-developed system we would expect the user to be given some kind of feedback on where their attempted emphasis fell.

The sentences are split into five primary sections, labelled **A** to **E** for the purposes of this report. The sections are presented to each subject in the same order (avoiding learning bias over the session), but sentences *within* each section are presented in different orders according to various Latin Squares (where the size of the square depends on the number of sentences in the section). To illustrate by example, section **B** contains two sentences (B1 and B2), each with two words to be emphasised on separate attempts (-a and -b). Thus the Latin Square appears as follows:

Table 5.1: *Example Latin Square design*

User	1st sentence	2nd sentence	3rd sentence	4th sentence
1	B1-a	B1-b	B2-a	B2-b
2	B2-a	B2-b	B1-a	B1-b
3	B1-b	B1-a	B2-b	B2-a
4	B2-b	B2-a	B1-b	B1-a

Since sections **B**-**E** all contained 2, 3, or 4 sentence options, and tests were carried out on 12 users, we could ensure that each sentence order was played to the same number of users (6, 4, or 3) for these sections, using 2x2, 3x3 and 4x4 Latin Squares. Section **A** contains 5 sentences, so the sentence orders presented within this section were slightly unbalanced by necessity.

5.1.2 Sentence design

For clarity, sentence design is outlined for each section of the test alongside the results (Section 6.1.1). A full list of all sentences used is included in Appendix B.

5.1.3 Pilot tests

Two pilot tests were carried out prior to finalising the experiment described. The first required the user to synthesise a number of different sentences generated by the author (different to those used within the final test), to assess how easy or hard someone with no experience found the task. An overall accuracy of around 50% was observed, and the user was found to improve markedly over the course of the session. Overall this confirmed that the type of task planned was suitable, and provided insight into the types of sentences that users may find easier/harder.

The second pilot test was of a similar structure to that of the final experiment design (outlined above). This pilot showed that although the main format for the test worked, it was too long, taking around 75 minutes in total. On the basis of this, the number of repetitions of each sentence was reduced from 10 to 8 (no learning curve was observed after 8 repetitions of a sentence), and two sections (not described here) were removed. Also, it was found that the user could not easily read the sentence script on the laptop screen, so a paper script was printed for use in the subsequent actual tests.

5.2 Listening test

5.2.1 Setup summary

In total, 33 subjects were recruited for a listening test, lasting 20-30 minutes depending on the subject. All subjects identified themselves as being native English speakers, and received £6 in compensation. Each user was presented with 92 sentences split across 7 sections, and asked to select one of two options in a 'forced choice' style test. This choice involved selecting either a preferred audio clip or a textual option, depending on the question.

Subjects completed the test under controlled conditions in the Perception Lab of the School of Informatics, Edinburgh[1]. Listeners were sat in a sound-proof both, listening to clips through headphones, played through a Google Chrome browser.

The experiment consisted of 8 pages, each containing between 8 and 16 questions. The sections were presented in the same order to each participant, but the specific questions selected within each section, and their ordering within the section, were randomly generated for each subject. Additionally, for each paired choice the order of the two options was randomised. This initial question selection and randomisation

[1] http://www.ppls.ed.ac.uk/staff/resources/experiment_booths.php

is performed using a Python CGI script on the first page of the test. The order of questions and choices to be displayed is written to a text file on the server with a unique user ID. As the subject progresses through the test, questions are loaded from this file (the user ID is passed through as an argument), and responses are written to a separate file (marked with the same ID) using PHP. Finally, post-experiment, a Python script checks each response file to ensure the correct number of questions have been answered. This flagged one case of a respondent clicking a 'submit' button twice, and the file was easy to correct.

Audio was embedded using HTML 5, in both *mp3* and *ogg* formats. The original sound files were recorded by routing the synthesis output through the Macbook Pro's headphone socket, an M-Audio USB interface and into Audacity running on a second machine. Files were recorded at 44.1kHz stereo and converted to *mp3* and *ogg* using dBpoweramp Audio Converter[2] at the highest quality possible. In total, 632 audio files have been created for the purposes of the experiment, of which around 180 are played to any one subject over the course of the experiment.

5.2.2 Sentence design

As per the generation test, a full list of all sentences is included in the Appendix B. For clarity, information discussing the types of sentences used in the test, and their generation, are discussed alongside the results, in Section 6.2. Note that when sentences have been said to have been randomly generated, the word lists are from the sources footnoted here[3].

5.2.3 Pilot tests

Two pilot tests were carried out on an initial version of the listening experiment. Respondents did not report any major issues with the experiment design, although the section investigating low accent interrogative semantics (described within results) was found to be 'difficult'. This is reflected in the results found overall. The pilots also revealed subjects completed the test more quickly than anticipated, which led to a small increase in the number of sentences presented in the actual experiment.

[2] http://www.dbpoweramp.com/

[3] http://www.speakenglish.co.uk/vocab/, http://www.babycentre.co.uk, http://www.languageguide.org/, http://www.myenglishpages.com

Chapter 6

Results and analysis

Chapter summary: *This chapter presents both quantitative and qualitative results from all stages of the generation and listening tests. The hypotheses tested were laid out in Section 4.*

6.1 Generation test

6.1.1 Quantitative results

Two-tailed binomial tests are used to mark 95% confidence intervals in tables and on charts. The mean of the binomial distribution is set to the proportion of correct emphases out of all attempts. Chi-squared tests are used to calculate p-values for significance when confidence intervals overlap. Accuracy is used as a primary metric, defined as the proportion of times that a user emphasises the intended word.

A - position of word in sentence:

Setup This section requires the user to emphasise five different words (contrastively) on separate runs of a sentence. The sentence is formed so that each word to emphasise is of two syllables, and is contained between words that cannot be emphasised by the system (i.e. function words). For example, *'We were HOPING to have a PARTY, so that PEOPLE can have a BOOGIE while they are HAPPY'*. In total, three such sentences exist - each user is presented with one of these.

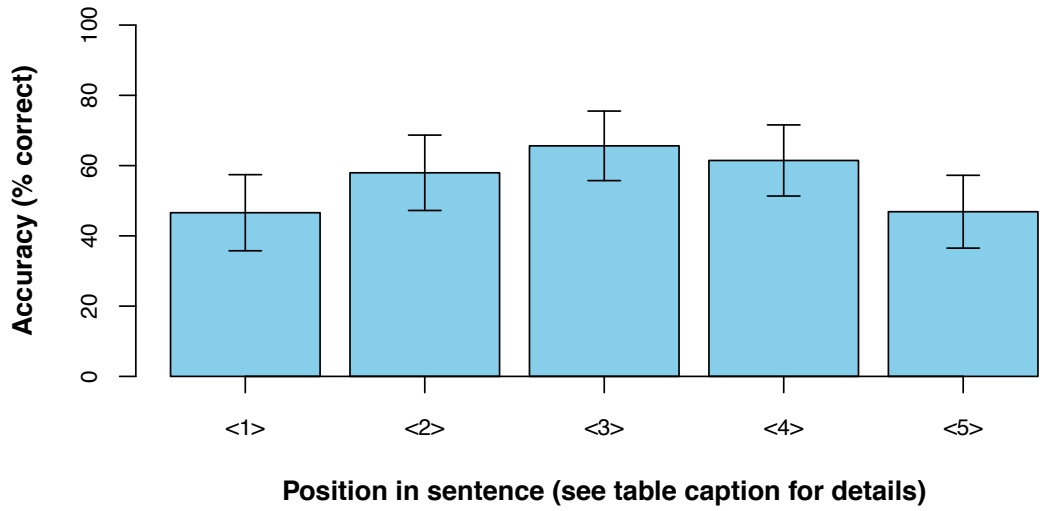
Results Results appear to show that users emphasise with the highest accuracy in the middle of a sentence, and less accurately towards the start and the end (Table 6.1,

Figure 6.1). It should be noted that only the difference between positions 1 & 3, and positions 3 & 5, are significant to $p < 0.05$. Users may find words at the start of the sentence more challenging to emphasise as they can be caught off-guard. There is no obvious reason why words towards the end of the sentence would be more difficult to emphasise. Most likely is that it is down to the rhythmic structure of the particular sentences used for this test, and not necessarily a general effect. Regardless of the driver, this does show that position in a sentence can have a small effect on how easy it is to emphasise a certain word, which is as hypothesised.

Table 6.1: *Effect of position of word in sentence, for example 'We were HOPING (1) to have a PARTY (2), so that PEOPLE (3) could have a BOOGIE (4) while they are HAPPY (5). Three such sentences exist, each user is assessed on one.*

Position in sentence:	**1**	**2**	**3**	**4**	**5**
Average accuracy:	47 ± 11%	58 ± 11%	66 ± 10%	61 ± 10%	47 ± 10%

Figure 6.1: *Effect of position of word in sentence*



B - naturalness of word to be emphasised:

Setup This section requires the user to emphasise two different words on separate runs of two different sentences (i.e. four separate runs in total). Within each sentence, emphasis would generally be considered appropriate on one of the words, whilst not on the other. To control for differences in difficulty according to the *position* of each

word in the sentence, the two sentences 'invert' the two words within their structure. This is best understood by considering the actual sentences:

- *'I used to like sausages, but it TENDS to be BREAD that now catches my imagination'*
- *'I used to like sausages, but it's BREAD that TENDS to catch my imagination now'*

Results Results are shown in Table 6.2. Although subjects performed slightly better when attempting to emphasis the appropriate word (*bread*) as opposed to an inappropriate word (*tends*), the difference is not significant ($p = 0.16$), and the null hypothesis (naturalness does not affect accuracy rate) cannot be rejected.

Table 6.2: *Effect of naturalness of word to emphasise*

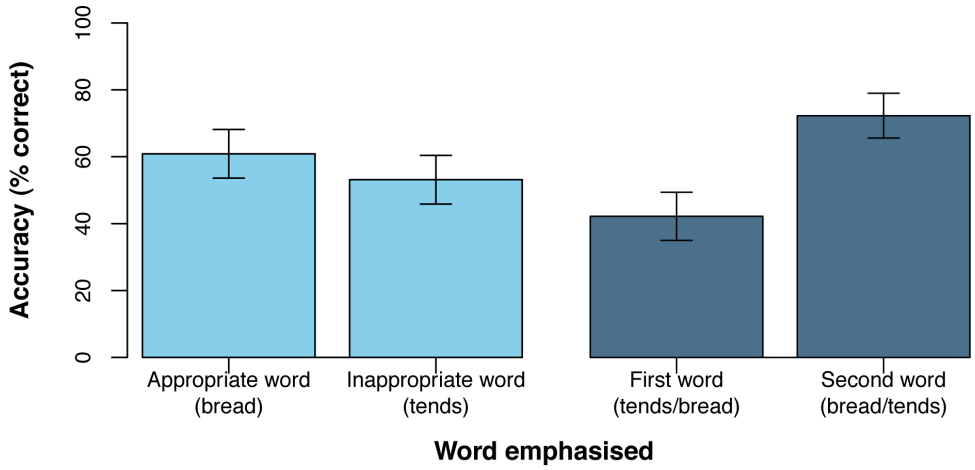
Word emphasised:	**Appropriate (bread)**	**Inappropriate (tends)**
Average accuracy:	$61 \pm 7\%$	$53 \pm 7\%$

Interestingly we can cut the data differently (Table 6.3, Figure 6.2) to compare the first word versus the second word in terms of position in the sentence (ignoring how appropriate the word is for emphasis).

Table 6.3: *Effect of position of word in the pair of sentences*

Word emphasised:	**First word** *(tends/bread)*	**Second word** *(bread/tends)*
Average accuracy:	$42 \pm 7\%$	$72 \pm 7\%$

Figure 6.2: *Naturalness of word / position of word (different cuts of same data)*



We see that emphasising the first word (whether this be *bread* or *tends*) has a significantly ($p = 7 \times 10^{-9}$) lower accuracy rate than emphasising the second. Users commented that this was a result of attempting to emphasise a word following a pause. The length of pause appears to be difficult for users to predict.

This result provides additional evidence to section **A** that the position in the sentence can affect ease of emphasis, when factors such as pauses and rhythm are taken into account.

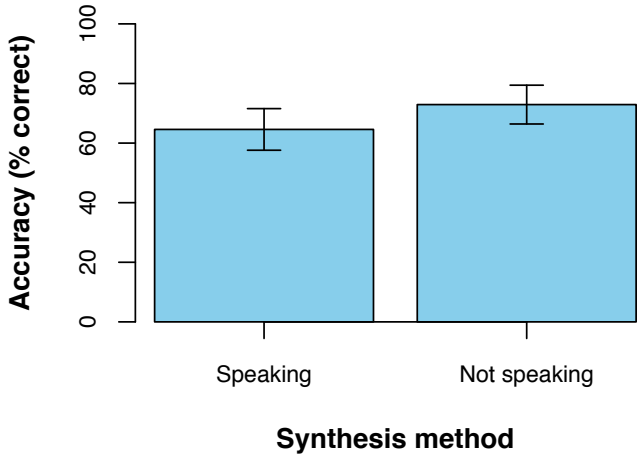
C - speaking alongside the synthesiser:

Setup This section requires the user to emphasise a two syllable word at the end of a sentence, not adjacent to any other content words. There are four sentences in total, two of which the user speaks out loud alongside the synthesiser whilst gesturing the emphasis, and two to emphasise without speaking, as per the rest of the test.

Table 6.4: *Effect of speaking alongside the synthesiser*

	Speaking alongside gestures	**Gestures only**
Average accuracy:	$65 \pm 7\%$	$73 \pm 7\%$

Figure 6.3: *Effect of speaking alongside the synthesiser*



Results Results show (Table 6.4, Figure 6.3) that speaking alongside the synthesiser results in a slightly lower accuracy level, though not to a significant level ($p = 0.10$). Thus the null hypothesis (speaking alongside the synthesiser does not

affect accuracy rate) cannot be rejected. Subjects tended to find the rhythm of the synthesiser different from their natural way of speaking, whilst some found it difficult to hear the synthesiser under their own voice, both of which may be reasons why accuracy does not improve.

D - emphasising two words within one sentence:

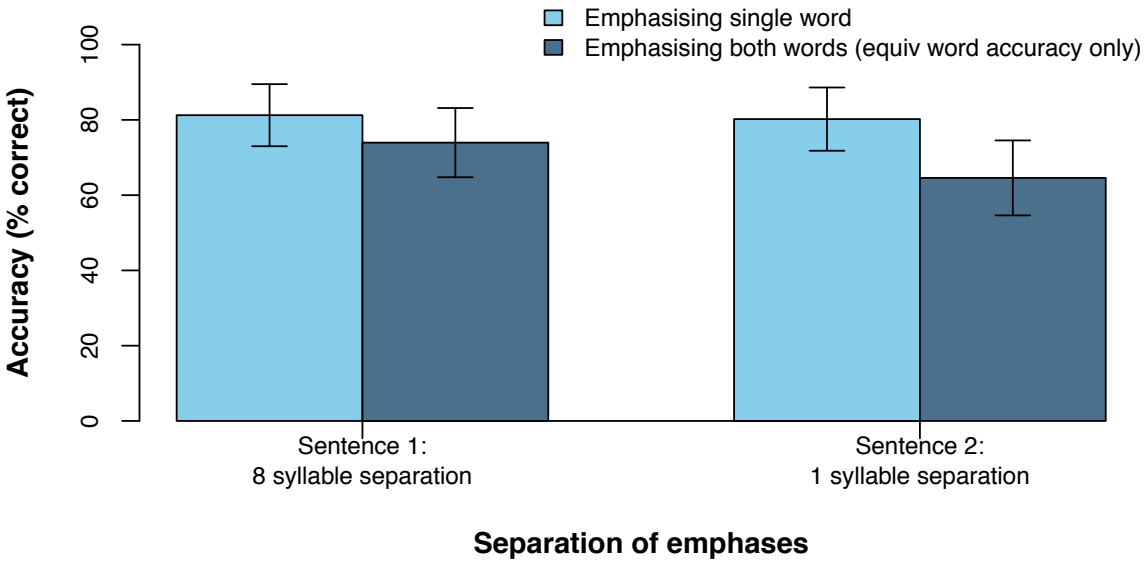
Setup This section requires the user to emphasise two words within the same sentence - one with general emphasis (right hand) and one with contrastive emphasis (left hand). Two sentences are presented, one in which the emphasised words are separated by one syllable (*'We planned it for THURSDAY, but FRIDAY was the day we ended up going'*), and one where they are separated by eight syllables (*'We planned it for THURSDAY, but we ended up going on FRIDAY'*). For both of these sentences, the user is also asked to emphasise just the second of the two words (i.e. *FRIDAY*) on separate attempts, as a control.

Results Results are shown in Table 6.5 and Figure 6.4. Directionally, it appears that emphasis accuracy decreases as extra words are emphasised within the same sentence. This result is as originally hypothesised. Furthermore, the closer the two words to be emphasised appear, the sharper the decrease in accuracy. However, the only statistically significant result is that of the decrease in accuracy in the case of the 1 syllable separation, marked in bold in Table 6.5 ($p = 0.02$).

Table 6.5: *Effect of emphasising two words per sentence (average accuracy), significant result in bold*

	Single word	**Double word** (equiv. word only)
8 syllable separation	$81 \pm 8\%$	$74 \pm 9\%$
1 syllable separation	$\mathbf{80 \pm 8\%}$	$\mathbf{65 \pm 10\%}$

Figure 6.4: *Effect of emphasising two words per sentence*



E - interrogative prosody with emphasis:

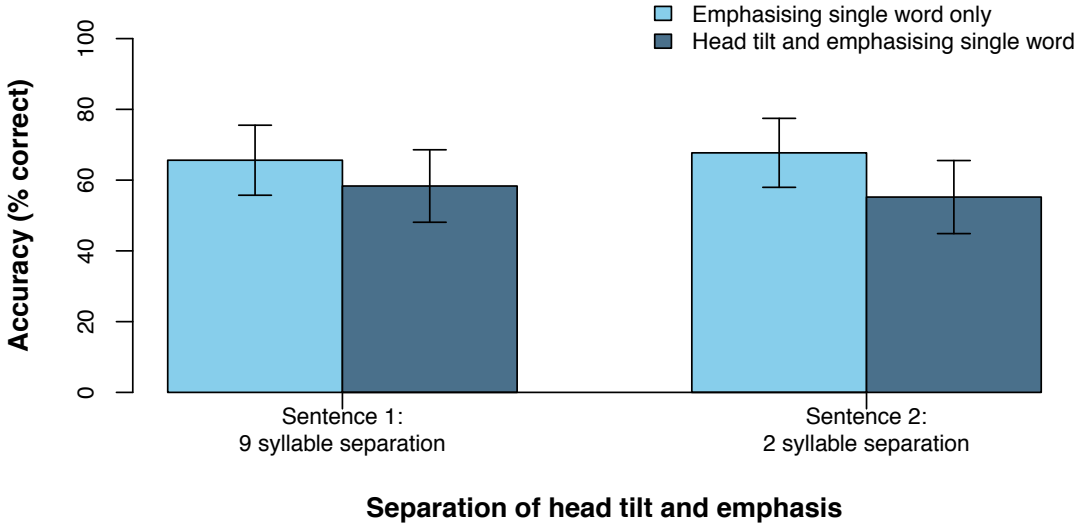
Setup This section requires the user to alter the prosody of two sentences to that of a WH-question (head-tilt to the left). In addition, the user must emphasise a specified word in the sentence (for example *'I'm here to renew my driving license. And what are YOU doing here today?'*). To observe the effect of the head tilt on accuracy, the user is also required to synthesise the same sentence with a plain contrastive emphasis gesture (without the head-tilt). In the same manner as in section **E**, the head-tilt and emphasis are separated by a small (2) and large (9) number of syllables on the two different sentences.

Results Results (Table 6.6, Figure 6.5) appear to be directionally similar to section **D** (emphasising two words in a sentence). The closer the head tilt and the emphasis, the lower the accuracy. However, no results are found to $p = 0.05$ due to the limited size of the data, thus we cannot reject the null hypothesis that adding a head-tilt in addition to the emphasis makes no difference to accuracy rate.

Table 6.6: *Effect of adding head tilt to emphasis (average accuracy)*

	Emphasis only	**Head tilt and emphasis**
9 syllable separation	$66 \pm 10\%$	$58 \pm 10\%$
2 syllable separation	$68 \pm 10\%$	$55 \pm 10\%$

Figure 6.5: *Effect of adding head tilt to emphasis*



Additional analysis - improvement over session:

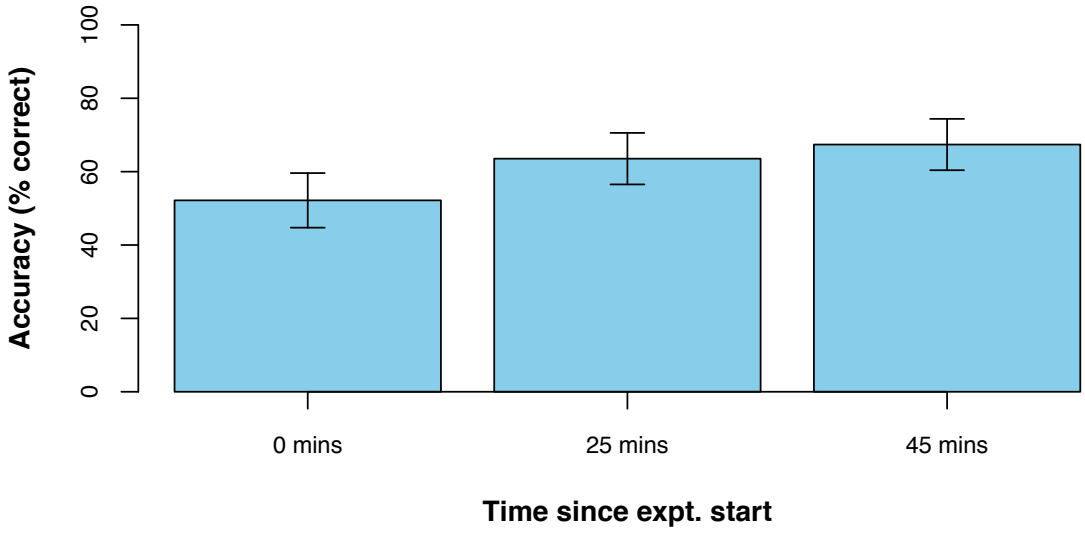
Setup Sentences from section **A** are re-used midway through the test (after section **C**), and at the conclusion (after section **E**). These are designed to determine any change in accuracy rate over the whole session.

Results Results show that users do improve over the course of the session, as hypothesised (Table 6.7, Figure 6.6). There is no significant difference between accuracy rates 25 minutes and 45 minutes into the test, though by 45 minutes users have improved to a significant level as compared to the start (recorded prior to training), with $p = 4 \times 10^{-3}$.

Table 6.7: *Improvement over session*

Time since experiment start:	**0 min**	**25 min**	**45 min**
Average accuracy:	$52 \pm 7\%$	$64 \pm 7\%$	$67 \pm 7\%$

Figure 6.6: *Improvement over session*



Additional analysis - spread of false negative and false positive gesture timings:

Setup Every attempt at emphasis made from sections **B** to **E** is amalgamated to assess the relative rate of early, correct and late emphases across the session. This is carried out both for *all* 8 attempts within each sentence, and then for the first attempt of each sentence only (which is ultimately the most important metric, as users are unlikely to want to repeat synthesis until they get the emphasis correct). As previously described, attempts are marked as:

- Very Early (**VE**) - emphasising wrong word
- Early (**E**) - no emphasis
- Correct (**C**) - correctly emphasised word
- Late (**L**) - no emphasis
- Very Late (**VL**) - emphasising wrong word

It should be noted that the sentences used across the session are designed to test certain aspects of the system, so are not necessarily representative of normal text. However, this should provide a directional result on the spread of emphasis attempts.

Results Firstly, averaging out over all 8 attempts for each sentence, the results are distributed as shown in Table 6.8. We can see that the correct emphasis is applied 64% of the time, emphasis is applied to an earlier word 6% of the time, and to a later word 6% of the time. The remaining 24% of the time, emphasis is applied slightly early or slightly late to the extent that the desired word is not emphasised, though no other words are erroneously emphasised either. Significantly more attempts are late than early, suggesting that if the reaction time of the system were made shorter, accuracy rates would improve.

Table 6.8: *Spread of emphasis gesture timings (8 attempts per sentence average)*

	VE	**E**	**C**	**L**	**VL**
Proportion of gestures:	6 ± 1%	5 ± 1%	65 ± 2%	18 ± 2%	6 ± 1%

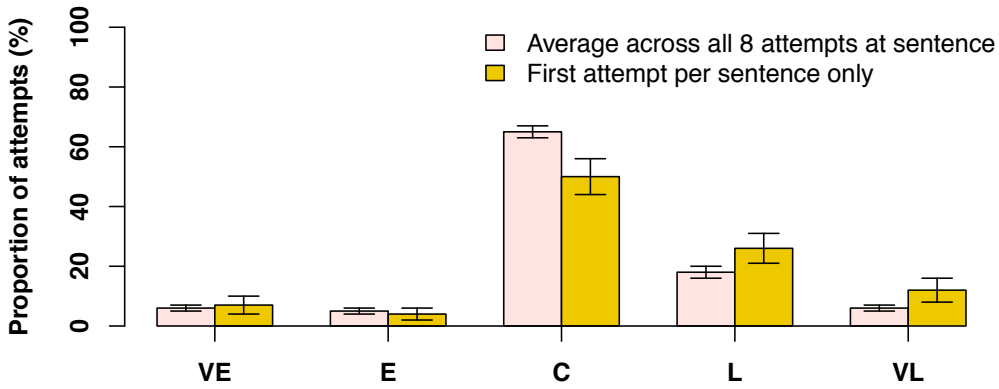
Turning to the *first* attempt at each sentence only (Table 6.9), we see that the correct emphasis is applied 50% of the time. Significantly more attempts result in no emphasis, or emphasising a subsequent word, compared to results averaged over all 8 attempts.

Although the percent correct (**C**) is lower on the first attempt, we can still reject the null hypothesis that the user emphasises words with an accuracy of chance (which would see far higher rates of **VE** and **VL**). Both the results averaged over 8 attempts, and for the first attempts, are laid out in Figure 6.7.

Table 6.9: *Spread of emphasis gesture timings (1st attempt only)*

	VE	**E**	**C**	**L**	**VL**
Proportion of gestures:	7 ± 3%	4 ± 2%	50 ± 6%	26 ± 5%	12 ± 4%

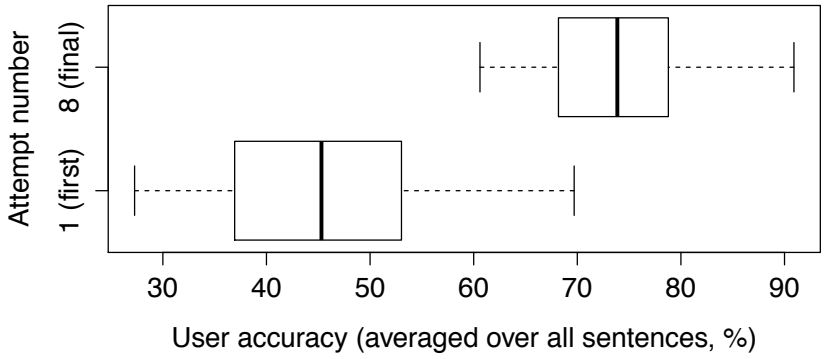
Figure 6.7: *Spread of emphasis gesture timings*



Additional analysis - user by user performance:

Setup It is interesting to look at users' success rates individually. Each user's success rate (accuracy) is tracked for both the first and final (eighth) attempts at each sentence synthesised during the test, and these results are averaged across sentences.

Figure 6.8: *Comparing first and eighth attempt of each sentence, averaged across all sentences, for each of the 12 users.*



Results A boxplot of results are shown in Figure 6.8. The range of accuracy rates for the 12 users is reasonably spread, particularly on the first attempt at each sentence. As expected, users improve significantly within the repetition of each sentence. This improvement is more marked than the gradual improvement over the course of the session seen previously. This shows that although most sentences are possible to emphasise correctly, the majority of users do benefit significantly from practice on the specific sentence (which is unlikely to be feasible in a real-life scenario).

6.1.2 Qualitative results

Users were also encouraged to provide verbal feedback on elements they found easy, hard, intuitive, unintuitive and so on. This section outlines recurring user comments, split by theme, and some general observations of the author.

Users' techniques: Users tended to adopt a variety of techniques to try and time the gestures appropriately. The main difference was in choosing when to initiate the rise of the beat gesture. There was an even split between those who tried to 'feel' the start of the gesture as naturally preceding the word to emphasise (as is intended by the author)

- and those who tried to anticipate when to lift their hand by counting the number of syllables before the word to emphasise. For example, in the sentence *'You need to put on a PLASTER'*, this latter set of users would learn that the hand needed to start rising just before the word *'to'*. However, in general it was those who tried to feel the motion as one smooth and natural gesture who appeared to perform better (though this has not been tracked quantitatively).

Other differences in technique included those who lifted their hand a long way (often to head-height), and those who tried to perform the gesture as efficiently as possible (barely raising the hand above waist-height). Some users were especially aggressive with their downstroke (which was justified by the claim that they felt more 'in control' by doing so) - to the extent that one user complained of a sore arm the following day.

Ease of use: Again, users provided a range of responses when it came to how 'easy' they found the system to use. One recurring theme is that users commented that the anticipation required to emphasise a word was slightly too long. This can also be seen in the quantitative results with the proportion of gestures registered as late (**L**) and very late (**VL**) being larger than those registered as early. This resulted in some users commenting that it felt too much like 'manipulating a machine' and not enough like 'performing a natural effect'. One user (with the highest overall accuracy rate) exclaimed 'Oh, too natural!' at one point, when performing a gesture slightly too late.

Conversely, a smaller number of users did comment that they felt as if they were emphasising the speech reasonably naturally, once their technique had been perfected.

Another recurring (and expected) theme was that of users finding the sentences in which two gestures had to be performed in close succession unnatural. In this case the gestures had to be overlapped - for example, beginning a left hand emphasis whilst the right hand is still beating in *'We planned it for THURSDAY, but FRIDAY was the day we ended up going'*. In everyday speech, we are unlikely to use alternate limbs in this way, and more likely to use the same limb for both emphases, or to only emphasise the more significant of the two words.

One final issue brought up by multiple users was in the difficulty of predicting the speed at which the synthesiser would utter certain words. Not knowing this, or how long the synthesiser would leave for pauses at commas, meant that users would often struggle to trigger the emphasis at the correct moment.

Naturalness of gestures: In general users found the beat gesture required to trigger the emphatic prosody to be a natural one. Two of the more inquisitive users were keen to experiment as to how well they could control the system without performing the downward part of the beat gesture (once discovering only the upward movement is required to trigger the emphasis). However, most other users agreed that bringing the hand down on the emphasised word felt natural enough to do (and helped with timing), in full knowledge that it was not actually required to trigger the emphasis.

One user commented that her everyday beat gestures consist of smaller circular wrist motions, so lifting her whole arm was not a particularly natural movement. She hypothesised that the extra time this gesture takes sometimes led her to gesture after the required window. Situations such as this would benefit from a machine learning based solution, which could learn customised gestures for each individual user.

Conversely, the head-tilt (as may have been expected) was not regarded as such a natural gesture by any of the users. It would be desirable to recognise either shoulder shrugs or some other more natural gesture in order to trigger question prosody, in any future version of this work.

Visual feedback: The experiment was set up so that users stood facing the Kinect, with the author placed in between the user and Kinect, with the laptop screen in view. The user was not explicitly introduced to the skeletal tracking view shown on the screen. However, having noticed the screen, two separate users claimed it was useful in helping them time gestures (particularly as version used in testing showed the skeleton 'flash' as the wrist passed above the hip to trigger emphasis). Incorporating visual feedback to the user in a more sophisticated way would be an interesting extension given future work in this area. For example, a live 'autocue' could highlight any words which the user emphasises, moving an arrow under the script as the synthesiser talks.

False positives and false negatives: Depending on the user's movement style, unintended gestures were sometimes recognised by the Kinect. These included users tilting their head excessively whilst performing emphasis, triggering question prosody. The number of times this happened was tracked for seven of the participants, occurring on 1.8% of intended gestures, on average. Two users' styles of emphasis involved 'bouncing' the hand after the fall, which sometimes led to a second emphasis as the wrist passed back above the hip. The number of occurrences was not tracked.

At other times, the Kinect did not recognise a movement despite the user making an appropriate gesture, due to poor lighting conditions, angles, or other unknown factors. This occurred on less than 0.4% of gestures attempted.

Body shape and Kinect position: It was noted that the Kinect responded to different body shapes in slightly different ways. Some neutral hand and hip positions differed, to the extent that the 'general importance' mode was triggered frequently by some users, and not by others. In addition, it was found that the relative position of hands and hips as viewed by the Kinect differed according to the vertical position of the Kinect relative to the user. The differences were not large, but ensuring the Kinect is always set at a specific position relative to the user should be ensured within any future work.

Overall enjoyment: Users tended to enjoy the experience, although fatigue often appeared to set in towards the end of the 50-minute session. Two users who performed towards the top end in terms of accuracy commented that it would be more interesting to synthesise more challenging sets of gestures, as the tasks provided were too simple. Experimenting with more complex gesture routines would be an interesting and enjoyable direction to explore.

6.2 Listening test

Throughout this section, unless otherwise stated, the null hypothesis assumes a listener chooses an option from the forced choice test with equal probability - i.e. the options are equivalent. Two-tailed binomial tests are used to calculate p-values (stated when significant). Additionally, 95% confidence intervals are provided for all data points.

Contrastive emphasis - naturalness

Setup Text is generated of the form *'No, Jess had **salmon** for her **breakfast** yesterday'*. Underlined words are randomly selected from appropriate word lists. Three versions of the sentence are then synthesised and recorded, two on which the author emphasises one of the words in bold (using gestural control), and one in which no words are emphasised (i.e. a standard, neutral pHTS baseline). Two alternative *questions* are written by hand, each one corresponding with one of the emphasised audio clips. These are each recorded in the author's voice using a Samson C01 microphone, at 44.1kHz stereo.

Presented to the subject is one of the question clips spoken by the author, followed by a pair of synthesised clips - one neutral, and one emphasised (either on the 'expected' word, or the 'unexpected' word). The subject's task is to pick which of the responses sound more **natural**.

Figure 6.9: *Screen-shot from the listening experiment.*

The setup is best understood through an example. The following show the 4 possible combinations for one response sentence. There are 15 such sentences, resulting in 60 possible question combinations. 20 were played to the subject. We would hypothesise the user to pick the option marked in red. Capitalised words represent emphasis.

Example 1

Author's voice: *Did Jess have trout for her breakfast yesterday?*

Synthesised option 1: *No, Jess had SALMON for her breakfast yesterday.*

Synthesised option 2: *No, Jess had salmon for her breakfast yesterday.*

Example 2

Author's voice: *Did Jess have trout for her breakfast yesterday?*

Synthesised option 1: *No, Jess had salmon for her breakfast yesterday.*

Synthesised option 2: *No, Jess had salmon for her BREAKFAST yesterday.*

Example 3

Author's voice: *Did Jess have salmon for her lunch yesterday?*

Synthesised option 1: *No, Jess had SALMON for her breakfast yesterday.*

Synthesised option 2: *No, Jess had salmon for her breakfast yesterday.*

Example 4

Author's voice: *Did Jess have salmon for her lunch yesterday?*

Synthesised option 1: *No, Jess had salmon for her breakfast yesterday.*

Synthesised option 2: *No, Jess had salmon for her BREAKFAST yesterday.*

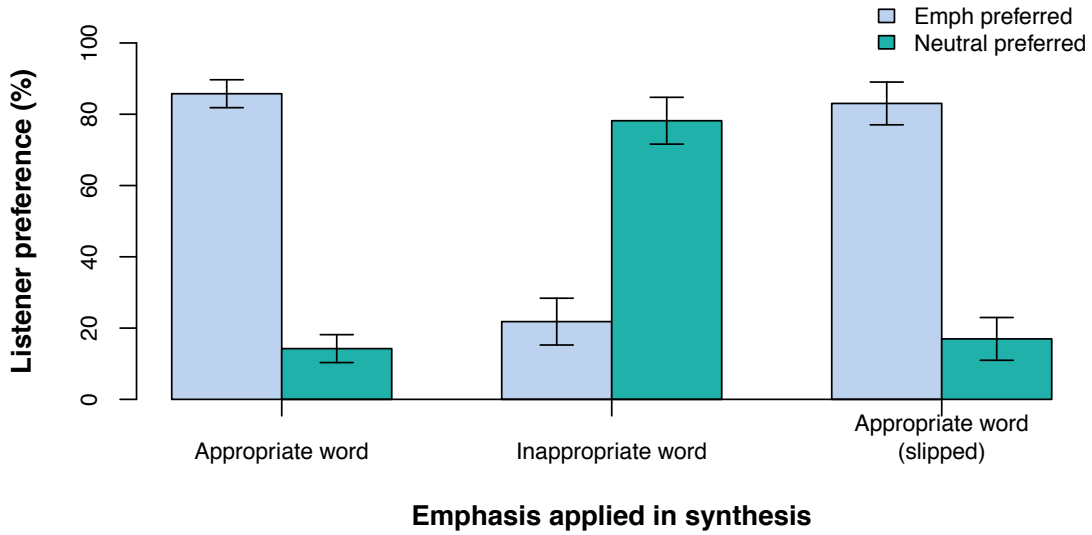
Results Results are shown in Table 6.10 and Figure 6.10. It has been found that listeners significantly prefer contrastive emphasis over neutral prosody when the em-

phasis is delivered on the appropriate word. Conversely, listeners significantly prefer the neutral prosody over emphasis delivered on an inappropriate word. Both of these results are as hypothesised.

Table 6.10: *Contrastive emphasis - naturalness*

Listener preference:	**Emphasised**	**Neutral**	***p-value***
Appropriately emphasised synthesis	**86** $\pm$ **4%**	14 $\pm$ 4%	$<$0.01
Inappropriately emphasised synthesis	22 $\pm$ 7%	**78** $\pm$ **7%**	$<$0.01
Appropriately emphasis (slipped) synthesis	**83** $\pm$ **6%**	17 $\pm$ 6%	$<$0.01

Figure 6.10: *Contrastive emphasis - naturalness*



Interestingly (and contrary to what was hypothesised), there is no significant difference between the performance of 'slipped' emphasis and correctly marked emphasis. 'Slipped' emphasis is defined as either emphasising the correct syllable half-way through (so that a slight glitch can be heard) or emphasising the word slightly late, so that although the remainder of the sentence is unaccented, the emphasised syllable is not raised. This is a positive result for the system - the user does not need to be 100% accurate with timing to improve the perceived naturalness.

Contrastive emphasis - semantics

Setup Text is generated of the form *'No, the **white** **dog** was **lying** on the surface'*. Underlined words are randomly selected from appropriate word lists to generate the

sentences. Different versions of the sentence are recorded, each with one of the bold words emphasised using the system created, or with neutral prosody to act as the pHTS baseline. 'Preceding' statements are then generated (in text form only), both that align appropriately with the emphasis, and that don't align. The subject is asked to select the initial textual statement that was **most likely** to have been said, given the synthesised response.

Again, this is best illustrated with an example. The following shows the 3 possible synthesised responses for one sentence. There are 15 such sentences in total, each recorded in 3 ways, resulting in 45 possible question combinations. 15 were played to the subject. We would hypothesise the user to pick the option marked in red (including for the user to pick randomly in the case of no emphasis in the response). Capitalised words represent emphasis. Further sentence examples can be found in Appendix B.

Example 1

Textual option 1: *The black dog was lying on the surface.*

Textual option 2: *The white mouse was lying on the surface.*

Synthesised audio: *No, the WHITE dog was lying on the surface.*

Example 2

Textual option 1: *The black dog was lying on the surface.*

Textual option 2: *The white mouse was lying on the surface.*

Synthesised audio: *No, the white DOG was lying on the surface.*

Example 3

Textual option 1: *The black dog was lying on the surface.*

Textual option 2: *The white mouse was lying on the surface.*

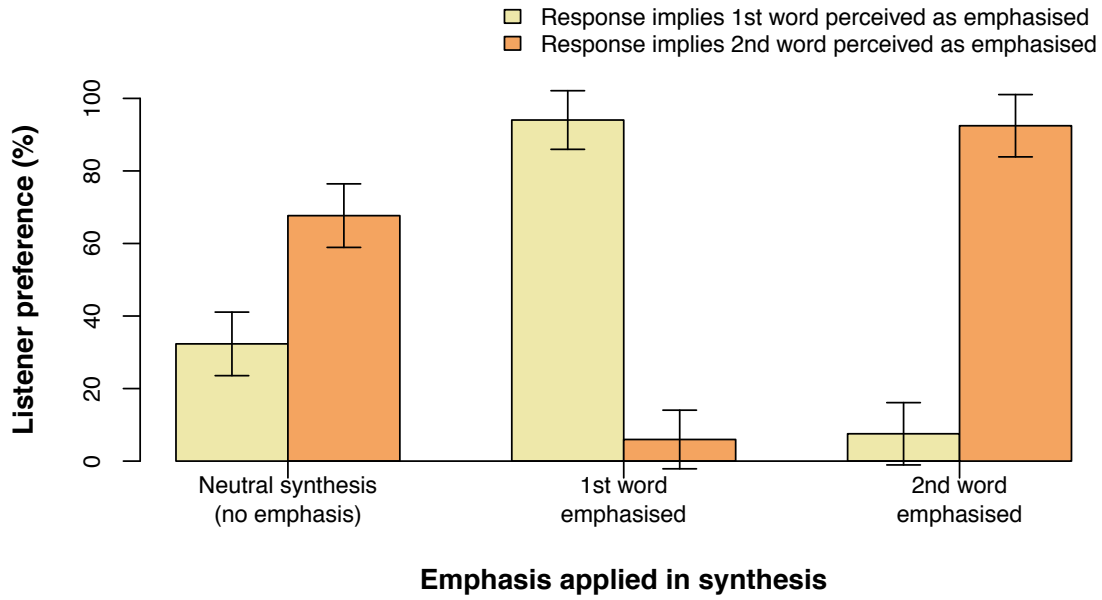
Synthesised audio (no emphasis): *No, the white dog was lying on the surface.*

Results Results (Table 6.11, Figure 6.11) show that adding contrastive emphasis does significantly alter the semantic interpretation of a sentence, as hypothesised. Users' choices for the most appropriate 'question' given an emphasised response correlate overwhelmingly with where the emphasis is placed in the response. The neutrally synthesised sentence shows a slight bias towards perceived emphasis in the latter parts of the sentence (i.e. 2nd word). Once one of the words is emphasised, the shift in perceived semantics is very significant.

Table 6.11: *Contrastive emphasis - semantics*

Choice indicates emphasis perceived to be on:	**1st word**	**2nd word**	***p-value***
Neutral synthesis	$32 \pm 7\%$	$\mathbf{67 \pm 7\%}$	<0.01
1st word emphasised in synthesis	$\mathbf{94 \pm 6\%}$	$4 \pm 6\%$	<0.01
2nd word emphasised in synthesis	$8 \pm 5\%$	$\mathbf{92 \pm 5\%}$	<0.01

Figure 6.11: *Contrastive emphasis - semantics*



Contrastive emphasis - position

Setup Four 1-syllable and four 2-syllable animal names were chosen from a word-list. These were placed into the two containers *'I didn't think it was an X, I thought it was a Y'* and *'I didn't think it was an X, I thought it was a Y or something'*. Of each of these containers, a neutral version and a version emphasising the Y animal were both synthesised. The subject was presented with 8 pairs of synthesised clips (one neutral, one emphasised in each case). The subject must choose which of the two sounds more **natural** for each pair. See Appendix B for the full list of sentences.

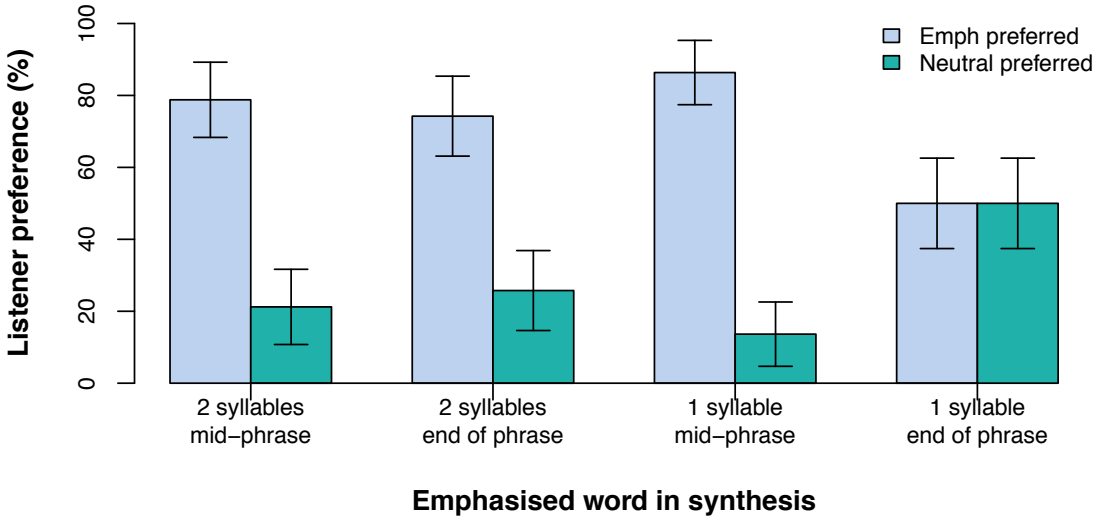
Results It has been found (Table 6.12, Figure 6.12) that contrastively emphasising the final syllable of a phrase does not increase perceived naturalness. However, emphasising the penultimate syllable (as is the case when a two syllable word is emphasised at the end of a sentence), or emphasising a yet earlier syllable, does improve

perceived naturalness. This result is statistically significant, and as hypothesised. The reason is likely to be that the accent currently implemented by the system (a simple pitch rise and duration increase) is unsuitable for the last syllable in a phrase, which as well as rising for emphasis, must fall during the syllable, to end on L%. As a result, the synthesis can sound unnatural. This is something that future iterations of the work must address, as emphasising the final syllable of a phrase is not uncommon.

Table 6.12: *Contrastive emphasis - position*

Listener preference:	**Emphasised**	**Neutral**	***p-value***
2 syllables, mid-phrase	**79 ± 10%**	21 ± 10%	<0.01
2 syllables, end of phrase	**74 ± 11%**	26 ± 11%	<0.01
1 syllable, mid-phrase	**86 ± 9%**	14 ± 9%	<0.01
1 syllable, end of phrase	50 ± 13%	50 ± 13%	

Figure 6.12: *Contrastive emphasis - position*



Low accent interrogative - naturalness

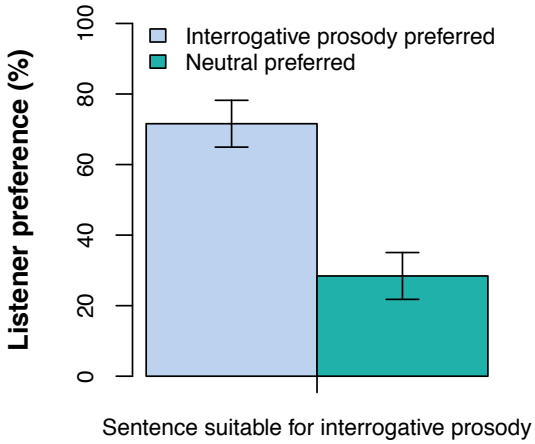
Setup 9 sentences were written by the author in which a low accent would sound appropriate. These are listed in Appendix B. These sentences were generated both with and without interrogative prosody (again, controlled using gestures on the developed system). Subjects were presented with the emphasised and neutral versions, and instructed to choose the one sounding more **natural**.

Results It has been found that an interrogative contour with low accent increases the perceived naturalness to listeners significantly (Table 6.13, Figure 6.13). This is as hypothesised. It should be noted that all sentences were designed to be appropriate to receive interrogative contour - this type of prosody has not been tested on sentences in which it would be *inappropriate* to receive the intonation, and therefore it cannot be said if (and by how much) naturalness would decrease when applied to inappropriate sentences. Future experiments should assess this scenario.

Table 6.13: *Low accent interrogative - naturalness*

Listener preference:	**Interrogative**	**Neutral**	***p-value***
Sentence suitable for interrogative prosody	**72** $\pm$ **7%**	28 $\pm$ 7%	$<$0.01

Figure 6.13: *Low accent interrogative - naturalness*



Low accent interrogative - semantics

Setup The setup is similar to that of the 'Contrastive emphasis meaning' test. 16 sentences (questions) have been written (by hand) in which a low accent interrogative prosody may be appropriate on two of the words in the sentence (see Appendix B). Three different versions are synthesised using the developed system - one neutral, and two with emphasis on different words. One of these is presented to the subject, followed by two textual options for feasible answers. The user is asked to choose the **most appropriate** answer, 'given the *way* in which the question was asked'. This is repeated 16 times, once per sentence. To once again illustrate by example:

Example 1

Synthesised audio: *So will you walk to John's HOUSE on Monday?*

Textual option 1: *No, I'll walk to his school.*

Textual option 2: *No, I'll drive to his house.*

Example 2

Synthesised audio: *So will you WALK to John's house on Monday?*

Textual option 1: *No, I'll walk to his school.*

Textual option 2: *No, I'll drive to his house.*

Example 3

Synthesis (no emphasis): *So will you walk to John's house on Monday?*

Textual option 1: *No, I'll walk to his school.*

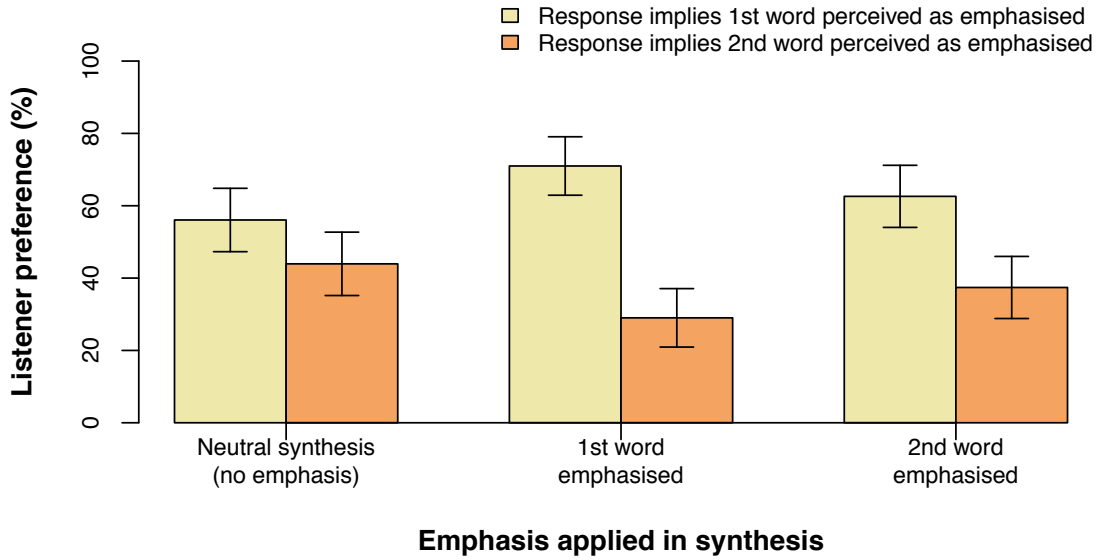
Textual option 2: *No, I'll drive to his house.*

Results Interesting results (Table 6.14, Figure 6.14) have been found when considering the semantics of synthesised speech with low accent interrogative prosody. As previously outlined, each of the two forced choice answers imply a different word in the synthesised text has been emphasised. In the neutral case, there is not a significant difference between those who answer having perceived the emphasis to be on the first of the words, and those who have perceived it be on the second of the words. However, placing a low accent on *either* of the two words leads people to answer as if the *first* of the two words have been emphasised. The effect is stronger when the first word is the one that is emphasised, though the effect is significant in the case of the second word being emphasised as well. Results are shown in Table 6.14. Although this has shown that the low accent interrogative contour does alter the perceived meaning, the result is not as hypothesised, as we would have expected emphasis on different words to alter the perceived meaning in *different* ways.

Table 6.14: *Low accent interrogative - semantics*

Choice indicates emphasis perceived to be on:	**1st word**	**2nd word**	***p-value***
Neutral synthesis	$56 \pm 9\%$	$44 \pm 9\%$	
Low accent on 1st word in synthesis	$\mathbf{71 \pm 8\%}$	$29 \pm 8\%$	<0.01
Low accent on 2nd word in synthesis	$\mathbf{63 \pm 9\%}$	$37 \pm 9\%$	<0.01

Figure 6.14: *Low accent interrogative - semantics*



There are a few factors which may lead to this outcome. Firstly, low accents are relatively rare in questions, as discussed within the literature review. A low accent tends to assume the subject is already active in the discourse. However, the sentences appearing in the text are removed from discourse. Secondly, the low accent effect is subtle, and by hearing the rising 'question' intonation at the end of the sentence, people appear to assume that the accented word is the first of the two, regardless of whether the accent comes. This may be as the first option for the emphasised word often has few alternatives, whereas the second option has many more. To illustrate with an example, for the sentence *'So will we drive to Lily's house later?'*, listeners are more likely to select *'No, we'll take the bus.'* than *'No, we'll drive to Kate's house.'* regardless of whether the low accent is on *drive* or *Lily's*. This may be because - without any context - there are fewer alternatives to *drive* (walk, bus, train) than there are to *Lily* (Kate, Sam, Gandalf, David Beckham etc...).

Low accent interrogative - question identification

Setup 24 sentences were generated using appropriate word-lists: 8 statements that could be interpreted as questions if the intonation were appropriate (e.g. *'She put the HAM in the freezer(?)'*), 8 WH-questions suitable for rising boundary tone (e.g. *'HOW many grapes are in the freezer?'*) and 8 WH-questions suitable for a falling boundary tone (e.g. *'Why is there an APPLE in the oven?'*). Each of the 24 sentences

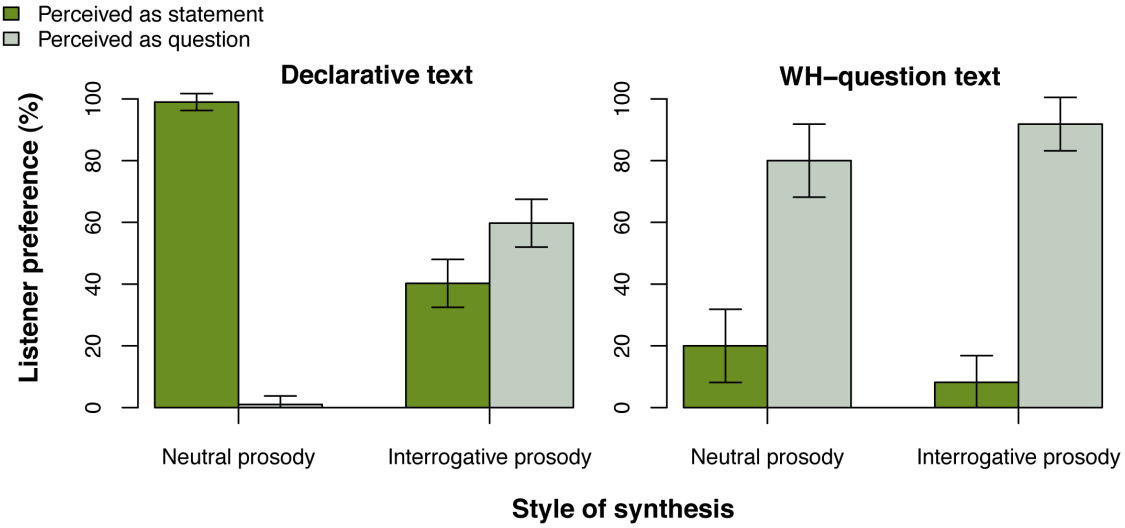
were synthesised with the appropriate interrogative prosody, and neutrally. The subject was presented with 14 forced choice pairs, and instructed to choose for each whether they believed the synthesis to be a statement or a question. This method has previously been used in [62] to assess the role of rising intonation in speech.

Results Results show (Table 6.15, Figure 6.15) that prosodic contours can convince a listener as to whether an ambiguous statement (for example *'She put the ham in the freezer'*) is a question or not. However, whatever the prosody, a WH-question is generally judged to be a question by listeners (i.e. neutral prosody does not convince most listeners that the utterance is a statement). This result is as hypothesised.

Table 6.15: *Low accent interrogative - question identification. Scores indicate proportion of users identifying utterance as a statement or a question*

	Statement	**Question**	***p-value***
Declarative text, neutral synthesis	**99 ± 3%**	1 ± 3%	<0.01
Declarative text, interrogative synthesis	40 ± 8%	**60 ± 8%**	<0.05
WH-text, neutral synthesis	20 ± 12%	**80 ± 12%**	<0.01
WH-text, interrogative synthesis	8 ± 9%	**92 ± 9%**	<0.01

Figure 6.15: *Low accent interrogative - question identification*



Thus for declarative statements, a synthesis system such as this has the advantage over a more basic system, in that it is able to convince listeners that a question is being

asked. Considering WH-questions however, although we have seen that *naturalness* is increased by altering the prosody, the perception of whether the text is a question or not is not affected significantly by the prosody, so the system designed here has no advantage in that regard.

Contrastive emphasis - accent height

Setup Finally, as described at the end of Section 4, a demonstrative test has been carried out to illustrate how parameter values may be fine-tuned by experimental means.

Five sentences have been written by hand, of the form *'I wanted to see X, but we're actually visiting **Y** on Tuesday'*, where *X* and ***Y*** are common names, ***Y*** is emphasised contrastively, and the day of the week is varied. Each sentence is synthesised with a contrastive pitch accent of +20%, using the system developed. Each of the syntheses are then modified in Praat, altering the pitch accent to values of +10%, +30% and +40%, in addition to the original +20%. A neutral version is also synthesised, referred to here as +0%.

The subject is presented with 10 forced choices. In each, one of the sentences compares two versions, each with a different height of pitch accent. For each subject it is ensured that each type of pitch accent is compared to each other one time (4+3+2+1=10 forced choices), and each of the five sentences is used twice. To illustrate with an example, two forced choices may be as follows:

Example 1

Option 1 - **pitch accent = 30%**:

I wanted to see Sam, but we're actually visiting ANDY on Monday.

Option 2 - **pitch accent = 10%**:

I wanted to see Sam, but we're actually visiting ANDY on Monday.

Example 2

Option 1 - **pitch accent = 0%**:

I wanted to see Emily, but we're actually visiting JESS on Sunday.

Option 2 - **pitch accent = 20%**:

I wanted to see Emily, but we're actually visiting JESS on Sunday.

Results Results show that a contrastive pitch accent increasing in pitch by between $\Delta 10$ and $\Delta 20\%$ is favoured by most listeners. This is slightly lower than originally designed by the author ($\Delta 25\%$) and used in the other tests previously outlined.

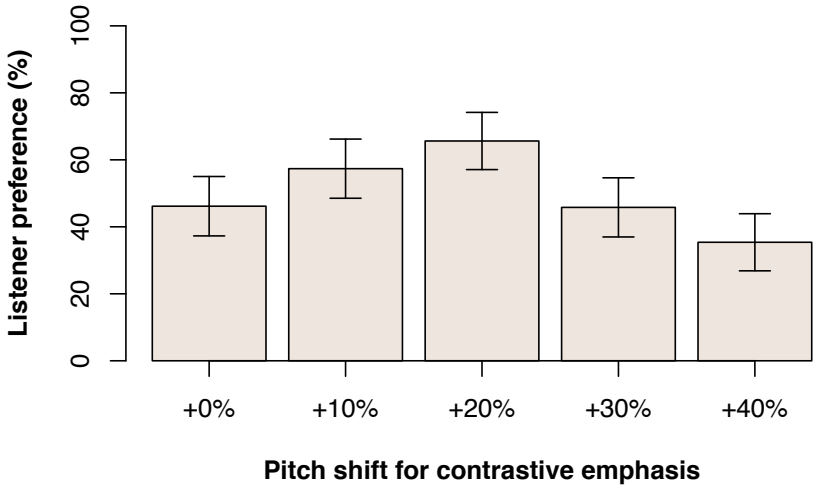
Table 6.16 shows direct preferences between pairs of pitch accent heights. Green indicates a pitch rise (row) gaining more than half of the 'votes' against a different pitch rise (column). Significant values are marked in bold.

Table 6.16: *Contrastive emphasis - pitch accent height. Green shading indicates the row label was preferred >50% of the time compared to the column label. Bold means the difference is significant to $p = 0.05$ over a null hypothesis of 50%.*

		Δ0%	Δ10%	Δ20%	Δ30%	Δ40%
Preferred pitch accent:	Δ0%		33%	41%	52%	59%
	Δ10%	**67%**		45%	58%	59%
	Δ20%	59%	55%		**66%**	**86%**
	Δ30%	48%	42%	34%		58%
	Δ40%	41%	41%	18%	42%	

An alternative way to look at the data is to consider only the proportion of all forced choices that each pitch accent value was preferred in. This does not take into account the alternative choice that each pitch accent is being compared with. The overall picture is similar however: Δ20% pitch accent is significantly preferred to all others apart from Δ10%, which performs similarly. The data in this form can be seen in Figure 6.16.

Figure 6.16: *Contrastive emphasis - pitch accent height*



Chapter 7

Discussion

Chapter summary: *This chapter discusses the results presented previously and how these tie into the initial problem statement. A short critical review is provided, before priority areas of focus for any future work are outlined.*

7.1 Summary and discussion of results

This project has aimed to develop a speech synthesis system that can be controlled in real time to improve the prosody of the output. Design choices made were based on previous work, preliminary experiments, and trial-and-error during development. The initial problem statement was split into six primary sub-questions, of which two have been studied in detail within this project, in the form of a generation test and a listening test.

7.1.1 Generation test

The generation test aimed to answer a number of questions on how well users can control the system, and how various factors affect accuracy rate. The original hypotheses and results are summarised and discussed here:

- **Overall accuracy rate over extended period of time**
  - ***Hypothesis:*** *users will emphasise correct words with a greater accuracy than that determined by 'chance'*
  - **Result:** supported by experiment, with users emphasising the correct word $\sim$50% of the time when controlling a sentence's prosody for the first time.

The wrong word is emphasised ∼20% of the time, and no word is emphasised in ∼30% of cases. This does suggest that the system is usable, and with improvements in latency, these percentages would be expected to improve further.[1]

- **Change of accuracy rate over course of session**
  - ***Hypothesis:*** *users will improve their accuracy rates over the course of a session*
  - **Result:** supported by experiment, with users improving significantly over the first 25 minutes. As an anecdotal aside, the author can generally emphasise with a significantly higher accuracy rate still, which suggests that long-term practice may improve accuracy rates further.
- **Repetition of a sentence**
  - ***Hypothesis:*** *users will improve their accuracy rate for a particular sentence with repetition*
  - **Result:** supported by experiment. However, this result is likely to be of little practical significance, as users would rarely wish to repeat the same sentence more than once.
- **Position of emphasised word within phrase**
  - ***Hypothesis:*** *the position in the sentence of the word to emphasise will affect a user's accuracy rate*
  - **Result:** supported by experiment, and hypothesised to be due to pauses and internal rhythms specific to individual sentences. This sort of discrepancy may be diminished as the pHTS engine is developed further, synthesising in a more rhythmically consistent manner.
- **Naturalness of word to emphasise**
  - ***Hypothesis:*** *accuracy rate will be higher for words that are more 'natural' to emphasise than words that wouldnt typically be emphasised in natural speech*

[1] It should be noted that the 'correct' emphasis itself can be split into 'correct - not slipped', and 'correct - slipped'. This was not detailed within the generation test's results section so as not to overcomplicate things. However, it should be noted that the correct emphasis was recorded as being 'slipped' ∼5% of the time in the generation tests. However, the listening test showed no perceptual difference in naturalness between 'slipped' and 'not slipped' synthesis, so the two are not considered separately anywhere else in this report.

  - **Result:** not shown to be a significant factor. This may show that users are experiencing the emphasis control as a somewhat 'abstract' task, without particularly linking the action to the meaning of the words being synthesised.

- **Speaking alongside the synthesiser**
  - ***Hypothesis:*** *accuracy rate will increase if a user speaks 'alongside' the synthesiser (due to increased awareness of position in sentence)*
  - **Result:** not shown to be true or a significant factor.

- **Number of gestures per sentence and their proximity**
  - ***Hypothesis:*** *accuracy rate will decrease if two gestures are performed in close proximity*
  - **Result:** supported by experiment in some cases. The fall in accuracy rate may be lessened if the latency of the system can be improved, as users would not have to 'overlap' two gestures that occur nearby within a sentence.

In summary, the generation test has shown that controlling a system such as this in real time is possible, indeed sometimes enjoyable, and that users' accuracy rates increase with practice. Various factors affect accuracy rate negatively, but most are related to latency issues in some form, and should therefore improve once buffering issues are resolved.

7.1.2 Listening test

The listening test aimed to answer whether perceived naturalness was improved by the prosodic modifications, and likewise if the semantic interpretation of synthesised phrases could be manipulated successfully through prosodic adjustments. These were tested using both high (H*) and low (L*) pitch accents, through contrastive emphasis and yes/no question prosody respectively. Additionally, the effect the position of emphasis in a sentence has on naturalness, question identification, and the optimal pitch shift parameter for contrastive emphasis were all explored. The original hypotheses and the results are summarised and discussed here:

- **Contrastive emphasis - naturalness:**
  - ***Hypotheses:*** *emphasis on correct word increases perceived naturalness, on incorrect word decreases perceived naturalness, slipped emphasis decreases perceived naturalness*
  - **Result:** the first of these two assertions were shown to be highly significant. However, the third (slipped emphasis decreases naturalness) was not supported by the experiment. Slipped contrastive emphasis also improves naturalness, despite the slight 'glitch' caused. This is positive for the purposes of our system, as the required gestural accuracy threshold is slightly lower.
- **Contrastive emphasis - semantics:**
  - ***Hypothesis:*** *correct emphasis can alter perceived semantics of sentence in an intended way, relative to neutral synthesis*
  - **Result:** supported by experiment, suggesting a system such as this will allow a user to make synthesis more expressive.
- **Contrastive emphasis - position in sentence:**
  - ***Hypotheses:*** *correct emphasis on final syllable of phrase decreases naturalness vs. neutral synthesis, emphasis elsewhere in sentence to increase naturalness vs. neutral synthesis*
  - **Result:** partly supported by experiment. Emphasis on the final syllable was not shown to be perceived as significantly more or less natural than neutral. All other emphasis positions were considered more natural. As previously mentioned, future work should investigate how to add realistic pitch accents that are suitable for the final syllable in a sentence (i.e. falling throughout the duration of a syllable).
- **Low accent interrogative - naturalness:**
  - ***Hypothesis:*** *interrogative synthesis with a low accent improves perceived naturalness vs. neutral synthesis*
  - **Result:** supported by experiment.
- **Low accent interrogative - semantics:**
  - ***Hypothesis:*** *position of low accent within interrogative prosody alters the perceived semantics of a sentence in an intended way, relative to neutral*

synthesis

  - **Result:** not supported by experiment. Although semantic interpretation was altered relative to a neutral utterance, emphasising *different* words didn't appear to change the interpretation.

- **Low accent interrogative - question vs. statement semantics:**
  - ***Hypothesis:*** *interrogative vs. neutral prosody affects the interpretation of declarative text, but a WH-word within the synthesised text overrides interrogative vs. neutral prosody*
  - **Result:** supported by experiment.
- **Contrastive accent pitch shift:**
  - ***Hypothesis:*** *some pitch peak (or peaks) are preferred by users significantly more than others*
  - **Result:** supported by experiment, with pitch peaks of +10% and +20% outperforming those of 0%, +30% and +40%. This suggests the value of +25% used for the contrastive emphasis peak in the rest of the test was too large.

In summary, the main takeaways are that naturalness is increased both in the case of contrastive emphasis using H*, and interrogative prosody using L*. However, representing prominence with a low pitch accent (L*) appears to be too subtle when it comes to altering the semantic content of a sentence, and so in future versions of this work it would be advisable to focus primarily on using the more common H* within any other prosodic effects created. These tests have also highlighted that finding a way to produce realistic pitch accents on the final syllable of a phrase is important.

It should be remembered that the other prosodic effects programmed - general emphasis, WH-questions, and general importance - have not been subject to listening tests. Given more time, these should be assessed in their own right, to ensure they also improve the quality of the synthesis as intended.

7.2 In the context of problem statement

Considered separately, results from the two main experiments within this project do suggest that it is possible to improve prosody in real time with gestural controls. The

issues that have occurred - primarily with regards to latency, and synthesising an overly subtle prosodic effect - are not areas that cannot be adequately addressed in future iterations of the work.

However, it must be noted that no 'cost-benefit' style analysis has been carried out regarding 'correct' vs. 'incorrect' emphases. For instance, currently around 50% of attempted first-time emphases are implemented correctly (as per the generation test), of which $\sim$90% may be found preferable by listeners when compared with neutral synthesis (as per the listening test). Meanwhile, on 20% of emphasis attempts the user may generate emphasis on the *wrong* word, which is perceived unfavourably relative to neutral synthesis $\sim$80% of the time. However, the 'cost' of an incorrect emphasis may still be larger than the 'benefit' of a correct emphasis (this has not been explicitly tested). Thus it should be tested if a passage that contains 5 words correctly emphasised and 2 words incorrectly emphasised is preferable to a neutrally synthesised version.

Other sub-questions, as discussed in Section 4, do also need to be considered as part of the 'pipeline' of questions making up the main problem statement. In summary however, the work carried out so far does appear to suggest that improved prosodic generation is definitely feasible using real time gestural control. It will be interesting to see how well a system such as this can perform with further improvements. Ultimately, a reasonable test for the system would take place in a less controlled scenario, where an expert user attempts to synthesise longer passages of natural conversation, with the output ultimately evaluated within a listening test (or by the subject with whom the conversation took place).

7.3 Critical review

Throughout, this report has outlined decisions made, their rationales, and the implications these decisions have had on the system made. However, for the sake of completeness a brief critical review is now included, outlining the areas in which compromises were made, and how these affected the outcome.

As previously discussed, rule-based prosody control has been implemented, as opposed to learning effects directly from data. On the scale of this project, this is likely to have led to more well defined prosodic effects, which have generally performed well in listening tests. However, in terms of scalability it would be preferable to learn parameters from data. This would allow a much more natural integration with the way that HMM-based synthesisers are already trained. This would ultimately lead to a more

flexible prosodic model, which may vary according to training data (different speakers may use subtly different prosodic effects).

Additionally, rule-based gesture recognition has been used in preference to machine learning techniques. For the purposes of this project, this has not affected the outcome significantly. In the long term however, machine learning methods must be implemented to allow the larger suite of gestures that would be required. Gestural and prosodic data would ideally be recorded simultaneously, with the system collecting data on the temporal relationship between the user's gestures and vocal effects.

Finally, it was not possible to successfully build the system on MAGE 2.0 in the time available, thus code based on MAGE 1.00 has been used. This has had an impact on the project in that the audio buffering affects the latency of the system, and therefore the accuracy rates possible. Future work should address this issue, as described below.

7.4 Future work

This section summarises the directions most likely to be priorities should this work be extended. Extensions are outlined in order of importance, with the most critical first.

As discussed, a primary issue in controlling the synthesiser accurately has been the relatively poor latency, which appears to be caused by audio buffering. Updating the code on which the project is based to that of MAGE 2.0 is likely to reduce the delay, as has been discussed previously. Upon improving the latency, many further doors will open in terms of accurate control. Firstly, more detailed algorithms may be used to recognise gesture timing more precisely, for instance using velocity and acceleration data of the hand to predict precisely where the peak of a beat gesture is likely to come relative to the synthesis. Secondly, more detailed prosodic rules to affect prosody *before* the user reaches the apex of their action may be implemented. For instance, immediately prior to a contrastive accent we may wish to decrease the speed of the speech slightly, as was shown to be the case in the preliminary experiments (Section 3.2).

Implementing the system's gesture recognition capabilities through machine learning methods will result in a much more flexible system going forward. New gestures for additional prosodic effects may be created more easily, gestures may be customised for those with accessibility requirements, and gestures may be classified more quickly from the set of known gestures. When using machine learning methods, thresholds may also be adjusted to optimise for precision and recall.

The prosodic effects used within this pilot have been chosen in terms of their hypothesised usefulness, ease of coding, and in order to demonstrate a variety of effects. More rigorous work should be conducted to determine a definitive set of prosodic effects that would be most useful for situations in which a device such as this may be used. In the long run, it would be more efficient and robust to learn these prosodic effects from data, rather than each being coded by hand. Additionally, more 'emotional' styles may be added to build a truly flexible and expressive system.

The emphasis prediction methods currently used (allowing pitch accents to be placed on any stressed syllable of a content word) work adequately, though are relatively rudimentary in nature. By incorporating a discourse model, the synthesiser may be able to assist the user in adding particular prosodic effects it predicts to be more likely - for instance biasing emphasis probabilities towards unusual words that are new to discourse. This would be implemented with the aim of improving users' accuracy rates further.

In summary, the system built for this project is in its early stages, but improvements such as those listed are both feasible and exciting, and should lead to a robust, novel and natural method of altering speech synthesis prosody in real time.

Appendix A

Flow charts describing gestural and prosodic rules implemented

Figure A.1: ***Contrastive emphasis:*** *Processes are iterated at every frame (15fps).*

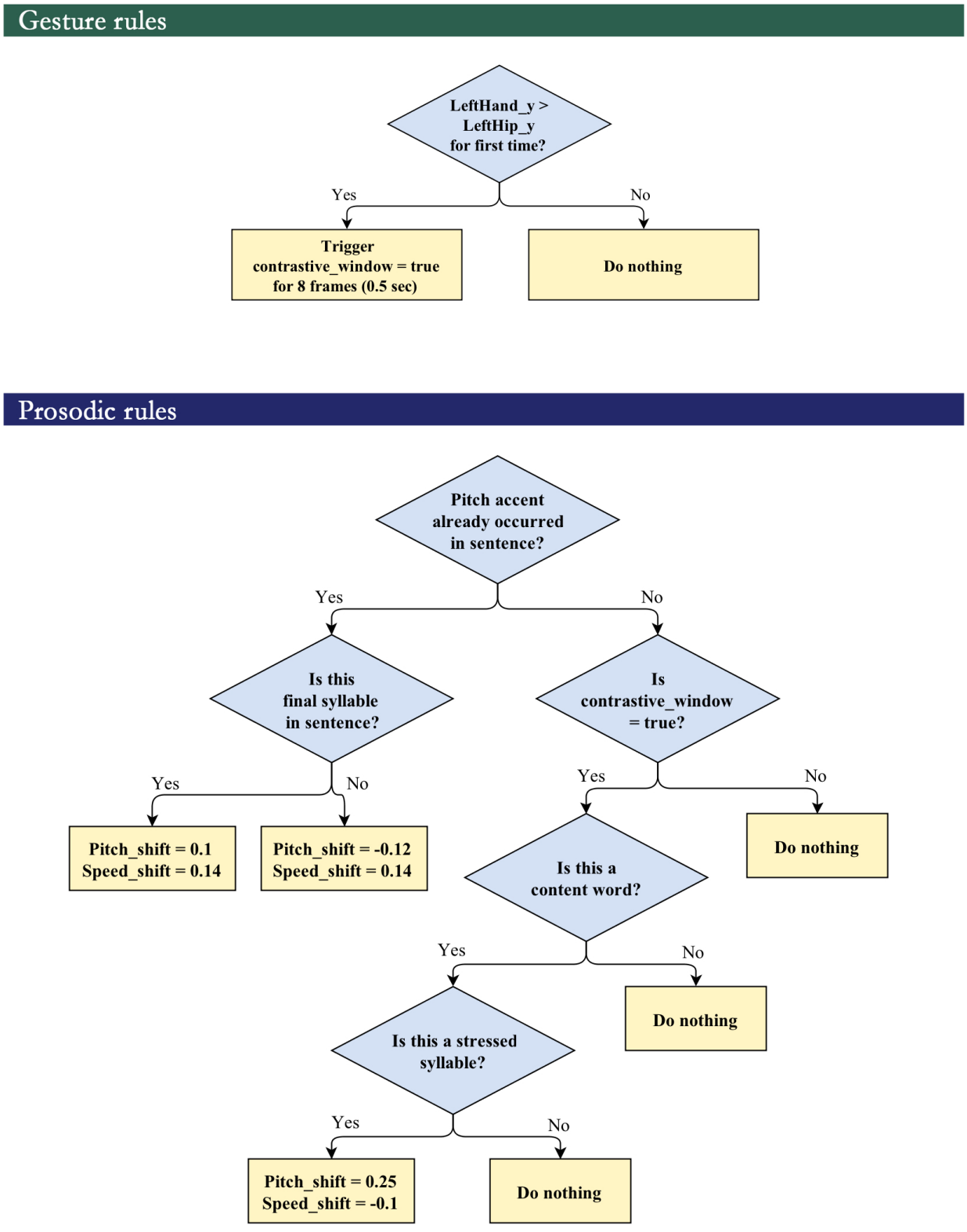


Figure A.2: ***General emphasis:*** *Processes are iterated at every frame (15fps).*

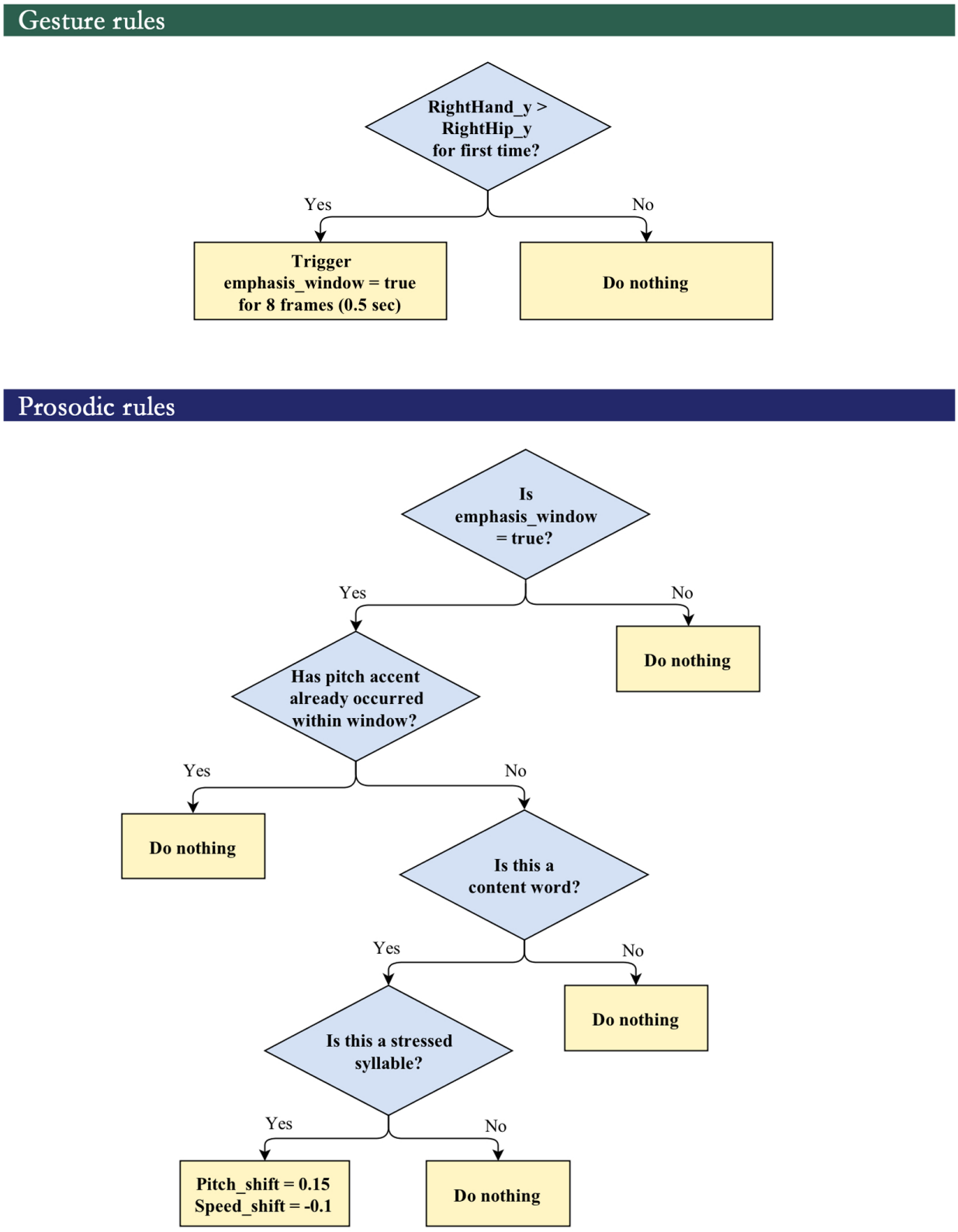


Figure A.3: ***Yes/No questions:*** *Processes are iterated at every frame (15fps).*

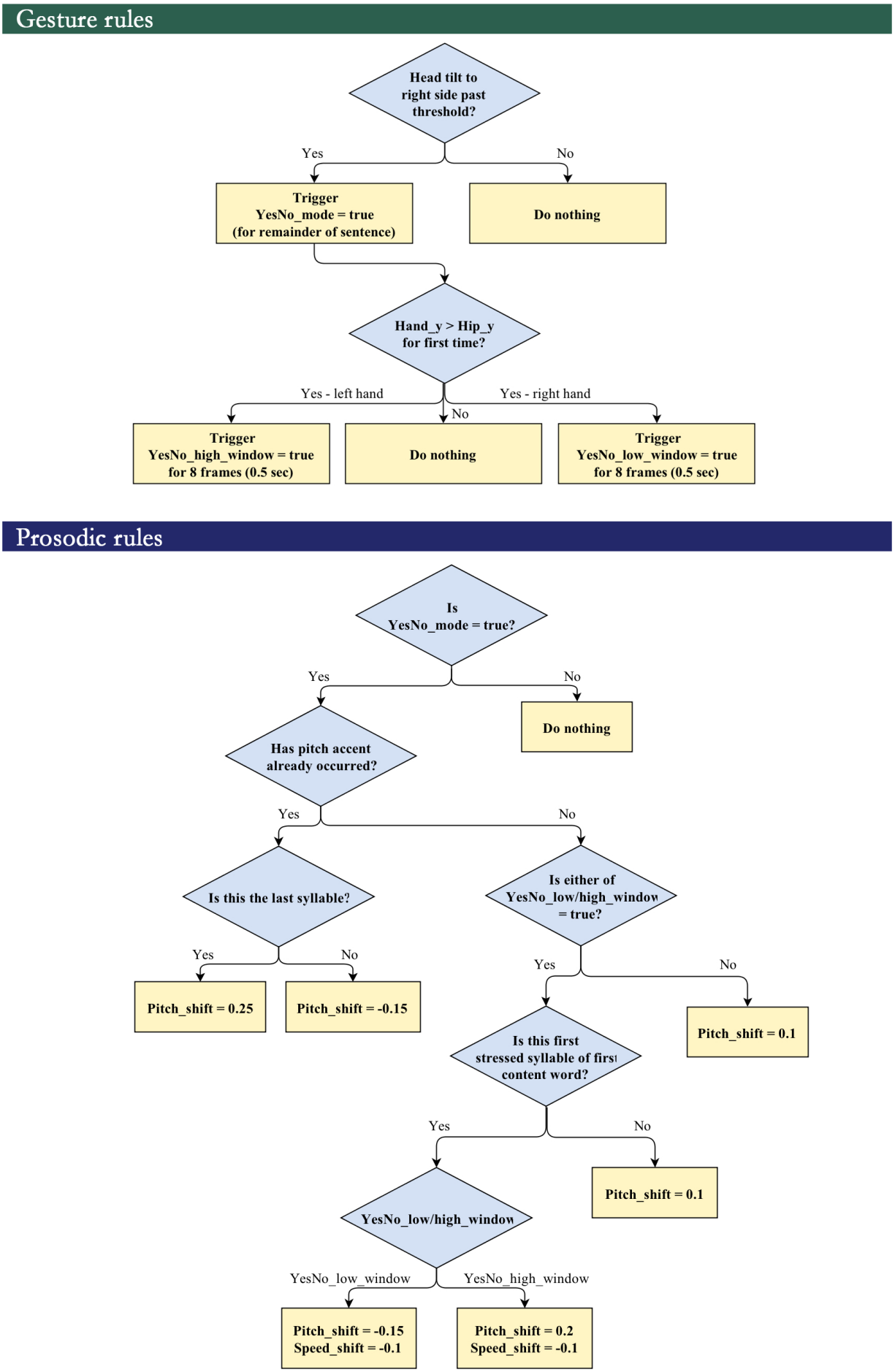
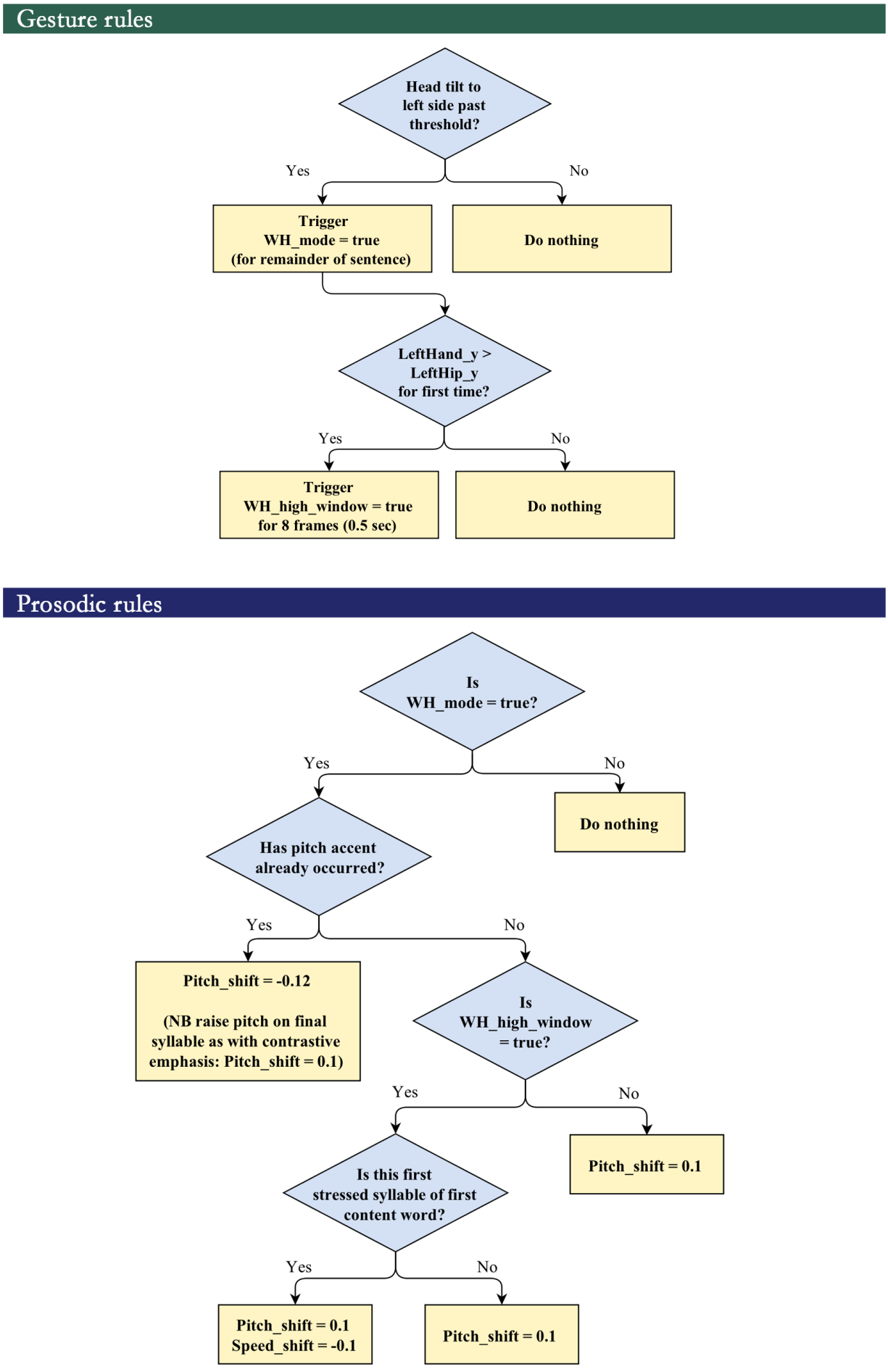


Figure A.4: ***WH questions:*** *Processes are iterated at every frame (15fps).*



Appendix B

Generation and listening test sentences

B.1 Generation test

Unless otherwise stated, words to emphasise are in capital letters, one emphasised per sentence.

Section A:

This is a REASON to have a CAMEL, so that PEOPLE can have a GIGGLE while it is EATING.

This is a REASON to have an APPLE, so that MONKEYS can have a NIBBLE while you are SLEEPING.

We were HOPING to have a PARTY, so that PEOPLE can have a BOOGIE while they are HAPPY.

Section B:

I used to like sausages, but it TENDS to be BREAD that now catches my imagination.

I used to like sausages, but it's BREAD that TENDS to catch my imagination now.

We were HOPING to have a PARTY, so that PEOPLE can have a BOOGIE while they are HAPPY.

Section C:

No dont use a bandage, you need to put on a PLASTER.

No I dont want a beer, Im in the mood for a WHISKY.

No its not a house, its what Id describe as a CASTLE.

No I dont care about being happy, what Im looking for is MONEY.

Section D: Two words to be emphasised where applicable.

We planned it for Thursday, but we ended up going on FRIDAY.

We planned it for THURSDAY, but we ended up going on FRIDAY.

We planned it for Thursday, but FRIDAY was the day we ended up going.

We planned it for THURSDAY, but FRIDAY was the day we ended up going.

Section E: With and without interrogative prosody (head-tilt).

Theres been a lot of talk about planning a holiday. But when are we actually going to TRAVEL to France?

Im here to renew my driving licence. And what are YOU doing here today?

B.2 Listening test

Contrastive emphasis - naturalness: Initial real-voiced questions in blue (one chosen per example), synthesised response in red (emphasis on one bold word, as discussed in Section 6.2).

Did they have pepper for their morning meal on Thursday?
Did they have radish for their evening meal on Thursday?
No, they had **radish** for their **morning** meal on Thursday.
Did Oliver have an apple for his lunch on Monday?
Did Oliver have a kiwi for his breakfast on Monday?
No, Oliver had a **kiwi** for his **lunch** on Monday.
Did I have chicken for my morning meal on Monday?
Did I have pork for my evening meal on Monday?
No, you had **pork** for your **morning** meal on Monday.
Did Sam have sandwiches for his lunch on Saturday?
Did Sam have toast for his breakfast on Saturday?
No, Sam had **toast** for his **lunch** on Saturday.
Did you have beans for your breakfast on Monday?
Did you have eggs for your supper on Monday?
No, I had **beans** for my **supper** on Monday
Did Jess have trout for her breakfast yesterday?
Did Jess have salmon for her lunch yesterday?
No, Jess had **salmon** for her **breakfast** yesterday.
Did you have pasta for lunch on Tuesday?
Did you have pizza for supper on Tuesday?
No, I had **pizza** for **lunch** on Tuesday.
Did they have crisps with their dinner yesterday?
Did they have cheese with their tea yesterday?
No, they had **cheese** with their **dinner** yesterday.
Did Olivia have bread for her lunch yesterday?
Did Olivia have pasta for her breakfast yesterday?
No, Olivia had **pasta** for her **lunch** yesterday.
Are they having blackcurrents for their tea on Sunday?
Are they having strawberries for their breakfast on Sunday?
No, they're having **strawberries** for their **tea** on Sunday.
Is Harry having steak for his dinner tomorrow?
Is Harry having muesli for his breakfast tomorrow?
No, Harry is having **muesli** for his **dinner** tomorrow.
Is Emily having turkey for her supper tonight?
Is Emily having chicken for dessert tonight?
No, Emily is having **chicken** for her **supper** tonight.
Did James have a pear for his tea on Thursday?
Did James have grapes for his breakfast on Thursday?
No, James had **grapes** for his **tea** on Thursday.
Did you have bacon for your supper on Tuesday?
Did you have beef for your lunch on Tuesday?
No, I had **beef** for my **supper** on Tuesday.
Did you have beans for your breakfast on Wednesday?
Did you have corn for your dinner on Wednesday?
No, I had **corn** for my **breakfast** on Wednesday.

Contrastive emphasis - semantics: Textual options in blue, synthesised response in red (emphasis on one of bold words, as discussed within Section 6.2).

The black dog was lying on the surface.
The white mouse was lying on the surface.
No, the **white dog** was lying on the surface.
The grey mouse was standing on the carpet.
The ginger cat was standing on the carpet.
No, the **ginger mouse** was standing on the carpet.
The brown cat was sitting on the chair.
The grey rabbit was sitting on the chair.
No, the **grey cat** was sitting on the chair.
The brown rabbit was sitting on the carpet.
The white dog was sitting on the carpet.
No, the **white rabbit** was sitting on the carpet.
The ginger rabbit was lying on the surface.
The grey dog was lying on the surface.
No, the **grey rabbit** was lying on the surface.
The grey dog was standing on the mat.
The grey rat was lying on the mat.
No, the grey **dog** was **lying** on the mat.
The black rat was lying on the table.
The black rabbit was standing on the table.
No, the black **rat** was **standing** on the table.
The ginger cat was lying on the chair.
The ginger rabbit was standing on the chair.
No, the ginger **rabbit** was **lying** on the chair.
The black dog was sitting on the chair.
The black cat was standing on the chair.
No, the black **cat** was **sitting** on the chair.
The white cat was sitting on the table.
The white dog was lying on the table.
No, the white **dog** was **sitting** on the table.
The ginger rabbit was standing on the surface.
The brown rabbit was lying on the surface.
No, the **ginger** rabbit was **lying** on the surface.
The white cat was standing on the carpet.
The black cat was sitting on the carpet.
No, the **white** cat was **sitting** on the carpet.
The brown rat was lying on the surface.
The black rat was standing on the surface.
No, the **black** rat was **lying** on the surface.
The grey rabbit was standing on the carpet.
The white rabbit was lying on the carpet.
No, the **white** rabbit was **standing** on the carpet.
The white mouse was lying on the mat.
The grey mouse was sitting on the mat.
No, the **grey** mouse was **lying** on the mat.

Contrastive emphasis - position in sentence: All sentences produced with and without emphasis on bold words.

I didn't think it was a bird, in fact, I thought it was a **cat**.
I didn't think it was a bird, in fact, I thought it was a **cat** or something.
I didn't think it was a cat, in fact, I thought it was a **bird**.
I didn't think it was a cat, in fact, I thought it was a **bird** or something.
I didn't think it was a rat, in fact, I thought it was a **mouse**.
I didn't think it was a rat, in fact, I thought it was a **mouse** or something.
I didn't think it was a mouse, in fact, I thought it was a **rat**.
I didn't think it was a mouse, in fact, I thought it was a **rat** or something.
I didn't think it was a leopard, in fact, I thought it was a **lion**.
I didn't think it was a leopard, in fact, I thought it was a **lion** or something.
I didn't think it was a lion, in fact, I thought it was a **leopard**.
I didn't think it was a lion, in fact, I thought it was a **leopard** or something.
I didn't think it was a hippo, in fact, I thought it was a **monkey**.
I didn't think it was a hippo, in fact, I thought it was a **monkey** or something.
I didn't think it was a monkey, in fact, I thought it was a **hippo**.
I didn't think it was a monkey, in fact, I thought it was a **hippo** or something.

Low accent interrogative - naturalness: All sentences produced with and without emphasis on bold words.

And is it only **Kate** that will come with us?
And was it only **Jack** that went to the football?
Is it only little **Martin** that cant tie his laces?
Do you **always** buy this many potatoes?
Do you **always** eat so much for your supper?
Is your dog **always** this happy when you get home?
Is your cat **normally** this moody when you feed her?
While we're in the queue, shall we get our **tickets** ready?
While were in the orchard, shall we pick some **apples** to take back with us?

Low accent interrogative - semantics: Textual options to pick in blue, initial synthesis in red, with one of the words in bold synthesised, as discussed within Section 6.2).

So will we make a **cake** for **Charlie** on Tuesday?
No, well buy it from the shop.
No, well make a cake for Daniel on Tuesday.
So will we buy a **cake** for **William** on Thursday?
No, we're going to make one.
No, we're only buying one for Sam.
So did you **make** the cake you gave to **Katie**?
No, I bought it actually.
No, I only made the one I gave to Sandra.
So will we **buy** a card for **Jack** this year?
No, we'll make a card for Jack.
No, we'll only buy one for James.
So will we **walk** to **Kates** house on Monday?
No, well drive to Kates on Monday.
No, well walk to Lilys house on Monday.
So will we **drive** to **Lily's** house later?
No, we'll take the bus.
No, we'll drive to Kate's house.
So will you **walk** to **Jon's** house on Monday?
No, I'll drive to his house.
No, I'll walk to his school.
So shall we catch the **train** to **Jack's** place tonight?
No, let's catch the bus.
No, let's get the train to Oliver's.
Is it only '**Little Harry**' that wont eat his greens?
No, 'Big Harry' wont eat them either.
No, 'Little Tommy' wont eat them either.
Is it only '**Big Sam**' who's allergic to nuts?
No, 'Little Sam' can't eat them either.
No, 'Big Mike' can't eat them either.
Is it only '**Fat Jon**' that likes sausages?
No, 'Skinny Jon' likes them too.
No, 'Fat James' likes them too.
Is it only '**Little Michael**' that can't tie his laces?
No, 'Big Michael' struggles with them too.
No, 'Little Zach' struggles with them too.
Did **Manchester City** beat **Arsenal** this season?
No, but Manchester United beat Arsenal twice.
No, but they did beat Chelsea twice.
Did **Spurs** beat **Chelsea** this season?
No, only Arsenal beat Chelsea this season.
No, but they did beat Charlton.
Did **Manchester United** beat **Everton** last season?
No, it was Manchester City that beat Everton.
No, but they did beat Liverpool.
Did **Everton** beat **Spurs** last season?
No, but Arsenal did.
No, but they beat Southampton.

Low accent interrogative - question vs. statement semantics: The first 8 sentences are declarative, the following 16 are wh-questions. Both are synthesised with neutral and interrogative prosody (emphasising words in bold). Note that when no

word is emphasised in a wh-question, the effect is that the wh- word itself is marked (i.e. the prosody remains flat until raising on the final syllable).

She put the sausages in the **freezer**(?)
She put the coffee in the **microwave**(?)
They put the carrots in the **cupboard**(?)
They put the peas in the **dishwasher**(?)
She put the **ham** in the freezer(?)
She put the **chicken** in the cupboard(?)
They put the **oranges** in the fridge(?)
They put the **blackberries** in the dishwasher(?)

Why did he put salami in the **cupboard**?
Why did she put salad in the **freezer**?
Why did they put the limes in the **microwave**?
Why did she put a turkey in the **refrigerator**?
Why is there an **apple** in the oven?
Why did they put the **tuna** in the cupboard?
Why did he put the **milk** in the freezer?
Why did he put **cabbage** in the fridge?
When did they put the cake in the oven?
When did she put the lemons in the cupboard?
How many grapes are in the freezer?
Why did he put the salmon in the microwave?
When did he put the sage in the cupboard?
How many pineapples are in the oven?
Why is there leek in the freezer?
Why is the kiwi in the cupboard?

Contrastive accent pitch shift: As described in Setion 6.2.

I wanted to see David, but we're actually
visiting **Josh** on Tuesday.
I wanted to see Rose, but we're actually
visiting **Lily** on Tuesday.
I wanted to see Emily, but we're actually
visiting **Jess** on Sunday.
I wanted to see Sam, but we're actually
visiting **Andy** on Monday.
I wanted to see Jacob, but we're actually
visiting **Russ** on Friday.

Bibliography

[1] Paul Taylor. *Text-to-speech synthesis*, volume 15. Cambridge University Press Cambridge, 2009.

[2] Cyril M Harris. A study of the building blocks in speech. *The Journal of the Acoustical Society of America*, 25:962, 1953.

[3] Dan Jurafsky and James H Martin. *Speech & Language Processing*. Pearson Education India, 2000.

[4] Simon King. Speech Synthesis course, Lecture and class notes (University of Edinburgh), 2013.

[5] Andrew J. Hunt and Alan W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *IEEE ICASSP-96*, volume 1, pages 373–376, Atlanta, GA, 1996. IEEE.

[6] Heiga Zen, Keiichi Tokuda, and Alan W Black. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064, 2009.

[7] Maria Astrinaki, Nicolas D'alessandro, Benjamin Picart, Thomas Drugman, and Thierry Dutoit. Reactive and continuous control of HMM-based speech synthesis. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 252–257. IEEE, 2012.

[8] Robert AJ Clark, Magdalena Anna Konkiewicz, Maria Astrinaki, and Junichi Yamagishi. Reactive control of expressive speech synthesis using Kinect skeleton tracking. *Information Processing Society of Japan*, 112(369):175–178, 2012.

[9] S Sidney Fels and Geoffrey E Hinton. Glove-talk: A neural network interface between a data-glove and a speech synthesizer. *Neural Networks, IEEE Transactions on*, 4(1):2–8, 1993.

[10] Robert Akl. *Evaluating appropriateness of EMG and flex sensors for classifying hand gestures*. PhD thesis, University of North Texas, 2012.

[11] Simon King. An introduction to statistical parametric speech synthesis. *Sadhana*, 36(5):837–852, 2011.

[12] Keiichi Tokuda, Heiga Zen, and Alan W Black. An HMM-based speech synthesis system applied to English. In *Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on*, pages 227–230. IEEE, 2002.

[13] Maria Astrinaki, Nicolas dAlessandro, and Thierry Dutoit. MAGE - A platform for tangible speech synthesis. In *Proceedings of the 12th Conference on New Interfaces for Musical Expression (NIME'12)*, 2012.

[14] Maria Astrinaki, Onur Babacan, Nicolas D'alessandro, Thierry Dutoit, and Sidney Fels. MAGE/pHTS in Collaborative Vocal Puppetry (CoVoP) - Multi-user performative HMM-based voice synthesis on distributed platforms. *QPSR of the numediart research program*, 4(3):59–64, 2011.

[15] Maria Astrinaki, Nicolas D'alessandro, and Thierry Dutoit. MageFaceOSC: Performative speech synthesis based on realtime face tracking. *QPSR of the numediart research program*, 5(1):15–16, 2012.

[16] Maria Astrinaki, Junichi Yamagishi, Simon King, Nicolas dAlessandro, and Thierry Dutoit. Reactive accent interpolation through an interactive map application. *Proceedings of the 14th Conference of the International Speech Communication Association (Interspeech 2013)*, 2013.

[17] Marc Schröder. Expressive speech synthesis: Past, present, and possible futures. In *Affective information processing*, pages 111–126. Springer, 2009.

[18] Janet E Cahn. The generation of a ect in synthesized speech. *Journal of the American Voice I/O Society*, 8:1–19, 1990.

[19] Iain R Murray and John L Arnott. Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication*, 16(4):369–390, 1995.

[20] Marc Schröder and Martine Grice. Expressing vocal effort in concatenative synthesis. In *Proc. 15th international conference of phonetic sciences*, pages 2589–2592, 2003.

[21] Takashi Nose and Takao Kobayashi. Recent development of HMM-based expressive speech synthesis and its applications. In *Proc. 2011 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2011)*, 2011.

[22] Akemi Iida and Nick Campbell. Speech database design for a concatenative text-to-speech synthesis system for individuals with communication disorders. *International Journal of Speech Technology*, 6(4):379–392, 2003.

[23] W Lewis Johnson, Shrikanth Narayanan, Richard Whitney, Rajat Das, Murtaza Bulut, and Catherine LaBore. Limited domain synthesis of expressive military speech for animated characters. In *Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on*, pages 163–166. IEEE, 2002.

[24] Volker Strom, Robert AJ Clark, and Simon King. Expressive prosody for unit-selection speech synthesis. In *INTERSPEECH*, 2006.

[25] Antoine Raux and Alan W Black. A unit selection approach to F0 modeling and its application to emphasis. In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pages 700–705. IEEE, 2003.

[26] Hiromichi Kawanami, Tsuyoshi Masuda, Tomoki Toda, and Kiyohiro Shikano. Designing speech database with prosodic variety for expressive TTS system. 2002.

[27] Volker Strom, Ani Nenkova, Robert AJ Clark, Yolanda Vazquez-Alvarez, Jason Brenier, Simon King, and Daniel Jurafsky. Modelling prominence and emphasis improves unit-selection synthesis. 2007.

[28] Wael Hamza, Ellen Eide, Raimo Bakis, Michael Picheny, and John F Pitrelli. The ibm expressive speech synthesis system. In *INTERSPEECH*, 2004.

[29] Junichi Yamagishi, Koji Onishi, Takashi Masuko, and Takao Kobayashi. Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis. *IEICE TRANSACTIONS on Information and Systems*, 88(3):502–509, 2005.

[30] Makoto Tachibana, Junichi Yamagishi, Takashi Masuko, and Takao Kobayashi. A style adaptation technique for speech synthesis using HSMM and suprasegmental features. *IEICE transactions on information and systems*, 89(3):1092–1099, 2006.

[31] Leonardo Badino, J Sebastian Andersson, Junichi Yamagishi, and Robert AJ Clark. Identification of contrast and its emphatic realization in HMM-based speech synthesis. 2009.

[32] Leonardo Badino, Robert AJ Clark, and Mirjam Wester. Towards hierarchical prosodic prominence generation in TTS synthesis. In *INTERSPEECH*, 2012.

[33] Kumiko Morizane, Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano. Emphasized speech synthesis based on hidden Markov models. In *Speech Database and Assessments, 2009 Oriental COCOSDA International Conference on*, pages 76–81. IEEE, 2009.

[34] Kai Yu, François Mairesse, and Steve Young. Word-level emphasis modelling in HMM-based speech synthesis. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4238–4241. IEEE, 2010.

[35] Yu Maeno, Takashi Nose, Takao Kobayashi, Yusuke Ijima, Hideharu Nakajima, Hideyuki Mizuno, and Osamu Yoshioka. HMM-based emphatic speech synthesis using unsupervised context labeling. In *INTERSPEECH*, pages 1849–1852, 2011.

[36] Peter Roach. A little encyclopaedia of phonetics. *University of Reading, UK*, 2002.

[37] Carlos Gussenhoven. *The phonology of tone and intonation*. Cambridge University Press, 2004.

[38] D Robert Ladd. *Intonational phonology*. Cambridge University Press, 2008.

[39] Gayle Marie Ayers. *Nuclear accent types and prominence: Some psycholinguistic experiments*. PhD thesis, Ohio State University., 1996.

[40] Janet Breckenridge Pierrehumbert. *The phonology and phonetics of English intonation*. PhD thesis, Massachusetts Institute of Technology. Dept. of Linguistics and Philosophy., 1980.

[41] Ann Wennerstrom. *The music of everyday speech: Prosody and discourse analysis*. Oxford University Press, 2001.

[42] Vivek Kumar Rangarajan Sridhar, Ani Nenkova, Shrikanth Narayanan, and Dan Jurafsky. Detecting prominence in conversational speech: pitch accent, givenness and focus. In *Proceedings of Speech Prosody*, volume 453, page 456. International Speech Communication Association Campinos, Brazil, 2008.

[43] Carla Umbach. On the notion of contrast in information structure and discourse structure. *Journal of Semantics*, 21(2):155–175, 2004.

[44] Nancy Hedberg. The prosody of contrastive topic and focus in spoken English. *Pre-proceedings of the workshop on information structure in context*, 14-52, 2002.

[45] Charles L Hamblin. Questions in Montague English. *Foundations of language*, 10(1):41–53, 1973.

[46] Knud Lambrecht and Laura A Michaelis. Sentence accent in information questions: Default and projection. *Linguistics and philosophy*, 21(5):477–544, 1998.

[47] Nancy Hedberg and Juan M Sosa. The prosody of questions in natural discourse. In *Speech Prosody 2002, International Conference*, 2002.

[48] Nancy Hedberg, Juan M Sosa, and Lorna Fadden. Tonal constituents and meanings of yes-no questions in American English. In *Proceedings of Speech Prosody*, 2006.

[49] Nancy Hedberg, Juan M Sosa, Emrah Görgülü, and Morgan Mameni. Prosody and meaning of wh-questions in American English. In *Speech Prosody 2010-Fifth International Conference*, 2010.

[50] Christine Bartels. *The intonation of English statements and questions: a compositional interpretation*. Psychology Press, 1999.

[51] Dwight Le Merton Bolinger. *Intonation and its uses: Melody in grammar and discourse*. Stanford University Press, 1989.

[52] Shuichi Nobe. Where do most spontaneous representational gestures actually occur with respect to speech? *Language and gesture*, pages 186–198, 2000.

[53] David McNeill. *Gesture and thought*. University of Chicago Press, 2008.

[54] Marc Swerts and Emiel Krahmer. The effects of visual beats on prosodic prominence. In *Proceedings of Speech Prosody*, 2006.

[55] Daniel P Loehr. *Gesture and intonation*. PhD thesis, Georgetown University, 2004.

[56] Thomas Leonard and Fred Cummins. The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, 26(10):1457–1471, 2011.

[57] Kam Lai, Janusz Konrad, and Prakash Ishwar. A gesture-driven computer interface using Kinect. In *Image Analysis and Interpretation (SSIAI), 2012 IEEE Southwest Symposium on*, pages 185–188. IEEE, 2012.

[58] KK Biswas and Saurav Kumar Basu. Gesture recognition using Microsoft Kinect. In *Automation, Robotics and Applications (ICARA), 2011 5th Intl. Conference on*. IEEE, 2011.

[59] Julia Hirschberg. Pitch accent in context predicting intonational prominence from text. *Artificial Intelligence*, 63(1):305–340, 1993.

[60] Heiga Zen. An example of context-dependent label format for HMM-based speech synthesis in English. *The HTS CMUARCTIC demo*, 2006.

[61] Maria Astrinaki, Nicolas D'alessandro, Loic Reboursière, Alexis Moinet, and Thierry Dutoit. MAGE 2.0: new features and its application in the development of a talking guitar. In *Proceedings of the 13th Conference on New Interfaces for Musical Expression (NIME'13)*, 2013.

[62] Ronald Geluykens. Intonation and speech act type: An experimental approach to rising intonation in queclaratives. *Journal of Pragmatics*, 11(4):483–494, 1987.